



HAL
open science

Optimal Sparsity Testing in Linear regression Model

Alexandra Carpentier, Nicolas Verzelen

► **To cite this version:**

Alexandra Carpentier, Nicolas Verzelen. Optimal Sparsity Testing in Linear regression Model. Bernoulli, 2021, 27 (2), pp.727-750. 10.3150/20-BEJ1224 . hal-02966853

HAL Id: hal-02966853

<https://hal.inrae.fr/hal-02966853>

Submitted on 18 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Optimal Sparsity Testing in Linear regression Model

Alexandra Carpentier and Nicolas Verzelen

April 24, 2020

Abstract

We consider the problem of sparsity testing in the high-dimensional linear regression model. The problem is to test whether the number of non-zero components (aka the sparsity) of the regression parameter θ^* is less than or equal to k_0 . We pinpoint the minimax separation distances for this problem, which amounts to quantifying how far a k_1 -sparse vector θ^* has to be from the set of k_0 -sparse vectors so that a test is able to reject the null hypothesis with high probability. Two scenarios are considered. In the independent scenario, the covariates are i.i.d. normally distributed and the noise level is known. In the general scenario, both the covariance matrix of the covariates and the noise level are unknown. Although the minimax separation distances differ in these two scenarios, both of them actually depend on k_0 and k_1 illustrating that for this composite-composite testing problem both the size of the null and of the alternative hypotheses play a key role.

1 Introduction

In the last decade, a lot of effort has been devoted to developing sound statistical methods for high-dimensional data. Most of the estimation procedures rely on the assumption that the parameter of interest has some possibly unknown structure. A prominent example is the high-dimensional linear regression problem where it is usually assumed that the regression parameter is sparse [7]. Despite the pervasiveness of the sparsity assumption in the literature, very few contributions challenge this assumption.

In this work, we tackle the largely ignored problem of assessing the sparsity of the regression parameter. Henceforth, we consider the random design high-dimensional linear regression model

$$Y = \mathbf{X}\theta^* + \sigma\epsilon, \quad (1)$$

where the unknown parameter θ^* belongs to \mathbb{R}^p , the noise vector $\epsilon \in \mathbb{R}^n$ follows a standard normal distribution and where the rows of \mathbf{X} are i.i.d. sampled according to the normal distribution $\mathcal{N}(0, \Sigma)$. For a given integer k_0 , we study the problem of testing whether the vector θ^* has at most k_0 non-zero components.

1.1 Minimax separation distance

Before discussing our contribution, we first formalize the sparsity testing problem. For a vector θ , $\|\theta\|_0$ denotes its number of non-zero entries. Then, given a non-negative integer $k_0 \in [0, p]$, write $\mathbb{B}_0[k_0] = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \leq k_0\}$ for the set of k_0 -sparse vectors θ . Rephrasing our aim, we want to test whether θ^* belongs to $\mathbb{B}_0[k_0]$.

In order to assess the quality of a testing procedure, we rely on the framework of minimax separation distances [29] which is described in the following paragraphs. Let $\|\cdot\|_2$ denote the l_2

distance in \mathbb{R}^p . For any $\theta^* \in \mathbb{R}^p$, $d_2(\theta^*, \mathbb{B}_0[k_0]) := \inf_{u \in \mathbb{B}_0[k_0]} \|\theta - u\|_2$ stands for its l_2 distance to the set of k_0 -sparse vectors. Intuitively, any α -level test ϕ of the null hypothesis $\{\theta^* \in \mathbb{B}_0[k_0]\}$ cannot reject the null with high probability when $d_2(\theta^*, \mathbb{B}_0[k_0])$ is too small. In this work, we aim at characterizing the smallest distance ρ , such that there exists a test achieving a small type I error probability and rejecting the null with high probability whenever $d_2(\theta^*, \mathbb{B}_0[k_0])$ is larger than $\rho\sigma$. These informal definitions are made precise in the next subsection. In the sequel, $\mathbb{P}_{\theta^*, \Sigma, \sigma}$ stands for the distribution of (Y, \mathbf{X}) in (1).

In high-dimensional linear regression, the intrinsic difficulty of estimation or testing problems sometimes depends on some specific features such as the knowledge of the noise level σ^2 or the knowledge of the distribution of the design. In this work, we focus on two emblematic settings. In the **independent** setting, we assume that the covariates are independent ($\Sigma = \mathbf{I}_p$) and the noise level σ is known. In the **general** setting, both the covariance of the covariates and the noise level are unknown.

1.1.1 Independent setting

Fix a positive integer $1 \leq \Delta \leq p - k_0$, we consider the alternative hypothesis where θ^* is $k_0 + \Delta$ -sparse. Given $\rho > 0$ and a test ϕ , we introduce its risk $R(\phi; k_0, \Delta, \rho)$ as the sum of the type I and type II error probabilities

$$R(\phi; k_0, \Delta, \rho) := \sup_{\theta^* \in \mathbb{B}_0[k_0]} \mathbb{P}_{\theta^*, \mathbf{I}_p, \sigma}[\phi = 1] + \sup_{\theta^* \in \mathbb{B}_0[k_0 + \Delta], d_2(\theta^*; \mathbb{B}_0[k_0]) \geq \rho\sigma} \mathbb{P}_{\theta^*, \mathbf{I}_p, \sigma}[\phi = 0], \quad (2)$$

where we only consider parameters θ^* in the alternative hypothesis that lie at a distances d_2 higher than $\rho\sigma$ from the null. For a fixed (known) $\gamma \in (0, 1)$, the separation distance $\rho_\gamma[\phi; k_0, \Delta]$ of ϕ is the largest ρ such that its risk is higher than γ , i.e. $\rho_\gamma(\phi; k_0, \Delta) := \sup\{\rho > 0 \mid R(\phi; k_0, \Delta, \rho) > \gamma\}$. Parameters θ^* lying at a distance larger than $\sigma\rho_\gamma(\phi; k_0, \Delta)$ from the null are therefore detected with probability higher than $1 - \gamma$ by ϕ . Finally, the minimax separation distance is

$$\rho_\gamma^*[k_0, \Delta] := \inf_{\phi} \rho_\gamma(\phi; k_0, \Delta), \quad (3)$$

where the infimum is taken over all tests ϕ .

1.1.2 General setting

In the general case, neither the covariance matrix Σ of the covariates, nor the noise level σ is known. We only assume that the the eigenvalues of Σ are bounded away from zero and from infinity. Respectively write $\eta_{\min}(\Sigma)$ and $\eta_{\max}(\Sigma)$ for its smallest and largest eigenvalues. Given $\eta > 1$, define

$$\mathcal{U}(\eta) = \{\Sigma : \eta^{-1} \leq \eta_{\min}(\Sigma) \leq \eta_{\max}(\Sigma) \leq \eta\}. \quad (4)$$

Fix $\rho > 0$. In this general model, the risk of a test ϕ is now taken as

$$\mathbf{R}_g(\phi; k_0, \Delta, \rho) := \sup_{\theta^* \in \mathbb{B}_0[k_0], \Sigma \in \mathcal{U}(\eta), \sigma > 0} \mathbb{P}_{\theta^*, \Sigma, \sigma}[\phi = 1] + \sup_{\sigma > 0, \theta^* \in \mathbb{B}_0[k_0 + \Delta], d_2(\theta^*; \mathbb{B}_0[k_0]) \geq \rho\sigma, \Sigma \in \mathcal{U}(\eta)} \mathbb{P}_{\theta^*, \Sigma, \sigma}[\phi = 0].$$

Since both Σ and σ are unknown, we evaluate the type I and type II error probabilities uniformly over all $\sigma > 0$ and all $\Sigma \in \mathcal{U}(\eta)$. The class of covariance matrices is constrained in $\mathcal{U}(\eta)$ in order to preclude too difficult settings where the eigenvalues of Σ differ too much to each other. Then, as in the previous subsection, the separation distance of a test ϕ is $\rho_{g, \gamma}(\phi; k_0, \Delta) :=$

$\sup \{\rho > 0 \mid \mathbf{R}_g(\phi; k_0, \Delta, \rho) > \gamma\}$ and the minimax separation distance in the general setting is defined by

$$\rho_{g,\gamma}^*[k_0, \Delta] := \inf_{\phi} \rho_{g,\gamma}(\phi; k_0, \Delta) . \quad (5)$$

In this work, we address both independent and general settings. More specifically,

- (i) We characterize the minimax separation distances in both the independent ($\rho_{\gamma}^*[k_0, \Delta]$) and the general ($\rho_{g,\gamma}^*[k_0, \Delta]$) settings by providing upper and lower bounds that match (up to a polylogarithmic loss in some regimes).
- (ii) We introduce computationally feasible testing procedures that (almost) simultaneously achieve this minimax separation distance over all Δ .

1.2 Previous results and related literature

Before further describing our contribution, we first discuss related results in the literature.

Signal detection. The signal detection problem which amounts to testing whether $\theta^* = 0$ is a special instance of the sparsity testing problem (corresponding to $k_0 = 0$). Signal detection in the Gaussian vector model (which corresponds to an orthogonal design) has been extensively studied [2, 18, 19, 23, 29] in the last fifteen years. More recently, this problem has also been investigated in the random design linear regression model [1, 16, 28].

To simplify the discussion, let us consider the high-dimensional setting where $p \geq n^{1+\zeta}$ for some fixed constant $\zeta > 0$. Then, one can deduce from [28] that the minimax separation distance in the independent setting satisfies

$$\rho_{\gamma}^{*2}[0, \Delta] \asymp_{\gamma,\zeta} \min \left[\frac{\Delta \log(p)}{n}, n^{-1/2} \right] ,$$

where $f(\Delta, n, p) \asymp_{\gamma,\zeta} g(\Delta, n, p)$ means that there exist positive constants $c_{\gamma,\zeta}$ and $c'_{\gamma,\zeta}$ (possibly depending on γ and ζ) such that $f(\Delta, n, p) \leq c_{\gamma,\zeta} g(\Delta, n, p) \leq c'_{\gamma,\zeta} f(\Delta, n, p)$ for all Δ, n , and p . For $\Delta \leq \sqrt{n}/\log(p)$, this separation distance is achieved by measuring the raw correlations between the response and the covariates and rejecting when too many of these correlations are unusually large. This can be done through the Higher-Criticism scheme [1, 28]. For denser alternatives ($\Delta \geq \sqrt{n}/\log(p)$), we start from the identity $\mathbb{E}Y_i^2 = \sigma^2 + \|\theta^*\|_2^2$ (and the Y_i are i.i.d.). Hence, a test rejecting when the empirical mean of Y_i^2 is significantly larger than σ^2 achieves the optimal squared separation distance of order $n^{-1/2}$ [1, 28]. In the specific regime where p is of the same order as n , and Δ is close to \sqrt{n} , the analysis has to be refined, see [16].

In the general setting (unknown Σ and unknown σ), it has been proved in [45] that,

$$\begin{aligned} \rho_{g,\gamma}^{*2}[0, \Delta] &\asymp_{\gamma,\xi} \frac{\Delta \log(p)}{n} && \text{if } \Delta \leq p^{1/2-\xi} \wedge \frac{n}{\log(p)} \text{ for any fixed } \xi \in (0, 1/2) ; \\ \rho_{g,\gamma}^{*2}[0, \Delta] &\geq c_{\gamma} \frac{\sqrt{p}}{n} && \text{if } \Delta \geq \sqrt{p} . \end{aligned}$$

However, for sparse alternatives, the corresponding test in [45] relies on a l_0 type variable selection method and has therefore exponential computational complexity. For denser alternatives ($\Delta \geq \sqrt{p}$), the lower bound entails that the minimax separation distance is large whenever $p \geq n^2$. Comparing both the independent and the general settings, we observe that the separation distance is significantly larger in the general setting for dense alternatives $\Delta \geq \sqrt{n}/\log(p)$.

Composite-composite testing problems and related work. An important difference between the signal detection ($k_0 = 0$) problem and the general sparsity testing problem ($k_0 > 0$) is that, in the latter, the null hypothesis is composite, thereby making the analysis of the problem more challenging. Up to our knowledge, the analysis of such composite problems has been considered only in a few work [3, 15, 20, 35], although the problems of constructing adaptive confidence regions (e.g. [8, 9, 12, 13, 27, 39]) or of functional estimation (e.g. [10, 11, 14, 25, 38]) are also related to such testing problems.

In particular, Nickl and Van de Geer [39] consider the problem of constructing adaptive and honest confidence sets for θ^* in the linear regression model (1) with known variance σ^2 . To achieve adaptivity to the unknown sparsity of θ^* , Nickl and van de Geer need to test hypotheses of the form $\|\theta^*\|_0 \leq k_0$. Following the so-called “infimum testing” principle, described in a systematic way in [26], they consider the statistic $\inf_{\theta \in \mathbb{B}_0[k_0]} \|Y - \mathbf{X}\theta\|_2^2/n$. This statistic corresponds to the infimum of the empirical variance when one corrects Y by a k_0 -sparse vector θ . Under the null, this statistic is not much larger than the noise level σ^2 . This leads them to derive

$$\rho_\gamma^{*2}[k_0, \Delta] \leq c_\gamma \left[n^{-1/2} + \frac{k_0 \log(p)}{n} \right],$$

for some $c_\gamma > 0$. Comparing this bound with its counterpart in the signal detection problem ($k_0 = 0$), we observe an increase by an additive term $\frac{k_0 \log(p)}{n}$ accounting for the complexity of the null hypothesis.

Up to our knowledge, it is still unknown whether the upper bound of Nickl and van de Geer is optimal (that is whether $\rho_\gamma^{*2}[k_0, \Delta]$ actually depends on $k_0 \log(p)/n$). In this manuscript, we answer this open question, this for all k_0 and Δ .

Sparsity testing in the Gaussian sequence model. The Gaussian sequence model $Y = \theta^* + \sigma\epsilon$ corresponds to case $p = n$ and $\mathbf{X} = \mathbf{I}_p$. In [17], we have pinpointed the minimax separation distances for all k_0 and Δ both when σ is known and σ is unknown. In particular, the optimal separation distance actually depends on the size k_0 of the null hypothesis for large k_0 but is significantly smaller than what is obtained by infimum tests strategies such as those in [26].

Generally speaking, [17] is closely related to the aims and results of this paper, but there is a significant challenge in adapting the results in [17] which are available for the Gaussian sequence setting, to the linear regression setting.

Related to this problem, some authors [10, 11, 33, 34] have considered the problem of estimating $\|\theta^*\|_0$ in the Gaussian sequence model in a Bayesian framework where all θ_i^* 's are sampled according to some mixture distribution. Although some of the ideas can be borrowed from their work, this Bayesian setting is quite different (see [17] for a discussion).

1.3 Our results

In this paper, we characterize the minimax separation distances $\rho_\gamma^*[k_0, \Delta]$ and $\rho_{g,\gamma}^{*2}[0, \Delta]$. To alleviate the discussion, we restrict ourselves throughout this paper to the high dimensional regime $p \geq n^{1+\zeta}$ where $\zeta > 0$ is an arbitrarily small absolute constant.

Independent setting. We establish matching (up to a multiplicative constants depending on γ) upper and lower bounds for $\rho_\gamma^*[k_0, \Delta]$, this for almost all values of k_0 and Δ ; see Table 1 for a summary of these results. An aggregated test is also shown to simultaneously achieve the optimal separation distance for all $\Delta > 0$, entailing that adaptation to the sparsity is possible for this

Table 1: Square minimax separation distances $\rho_\gamma^{*2}[k_0, \Delta]$ in the independent setting for $k_0 \in [1, p-1]$ and $\Delta \in [1, p-k_0]$ when $p \geq n^{1+\zeta}$ with a fixed $\zeta > 0$. Separation distances are given up to constants that may depend on γ and ζ .

k_0	Δ	$\rho_\gamma^{*2}[k_0, \Delta]$
$k_0 \leq p^{1/2-\zeta}$	$1 \leq \Delta \leq k_0 + \frac{\sqrt{n}}{\log(p)}$	$\frac{\Delta \log(p)}{n}$
	$k_0 + \frac{\sqrt{n}}{\log(p)} \leq \Delta \leq p - k_0$	$\frac{1}{\sqrt{n}} + \frac{k_0 \log(p)}{n}$
$p^{1/2+\zeta} \leq k_0 \leq c_\gamma \frac{n}{\log(p)}$	$1 \leq \Delta \leq k_0 p^{-\zeta}$	$\frac{\Delta \log(p)}{n}$
	$k_0 \leq \Delta \leq p - k_0$	$\frac{k_0}{n \log(p)}$

problem. In our exhaustive picture of $\rho_\gamma^*[k_0, \Delta]$, some of the regimes in k_0 and Δ are addressed by simple extensions of signal detection tests. However, other regimes turn out to be more challenging and require novel ideas. In what follows, we briefly mention these original aspects.

- We prove that, when $k_0 \geq cn/\log(p)$, then the testing problem becomes extremely difficult, in the sense that the separation distance $\rho_\gamma^*[k_0, \Delta]$ is very large. For $k_0 \geq n$, this separation distance is even infinite. This is not unexpected since identifiability problems arise in this regime.
- For moderate $k_0 \in [\frac{\sqrt{n}}{\log(p)}, p^{1/2-\zeta}]$ and large Δ , we prove that the upper bound of Nickl and Van de Geer [39] turns out to be optimal, i.e. the squared minimax separation distance is achieved by their infimum test and is of the order of $\frac{k_0 \log(p)}{n}$. The general idea is to reduce the problem of sparsity testing (with known variance) to a detection problem with unknown variance.
- For larger $k_0 \in [p^{1/2-\zeta}, cn/\log(p)]$ (where $\zeta > 0$ is an arbitrarily small absolute constant, and where $c > 0$ is an absolute constant), then both upper and lower bounds are new. The lower bound is based on moment matching strategies and best polynomial approximation akin to those of [17] in the Gaussian model. But the derivation is significantly more involved in the regression setting. For small Δ ($\Delta \leq k_0$), an optimal test is built using any estimator of θ^* achieving a small l_∞ error (see e.g. [31, 32, 44, 51]). The test simply rejects when this estimator has more than k_0 unusually large entries. For denser alternatives ($\Delta \geq k_0$), the approach is quite different. We build a statistic based on the empirical Fourier transform of some correction of the raw correlations between the covariates and the responses Y . This approach is reminiscent of sparsity estimators in [17, 33] in the Gaussian sequence model.

General setting. We derive lower and upper bounds of the minimax separation distance $\rho_{g,\gamma}^*[k_0, \Delta]$. These bounds match except in the large k_0 and Δ regime, where there is a $\log^2(p)$ mismatch. See Table 2 for a summary of the results. As in the independent setting, we emphasize below the most novel ingredient of our analysis.

- Achieving the optimal squared distance $\Delta \log(p)/n$ could be easily done if one has access to an estimator whose l_∞ distance to θ^* is less than $\sigma \sqrt{\log(p)/n}$ with high probability. However, such an estimator is unknown for general covariance matrices $\Sigma \in \mathcal{U}(\eta)$. For $\|\theta^*\|_0 \geq \sqrt{n}$, it is even proved that no such estimator exists [9]. Here, we first select a reasonable candidate

Table 2: Square minimax separation distances in the general setting (in the $\asymp_{\gamma, \eta}$ sense, see Sub-section 1.5). We report in this table only the case where $n^{1+\zeta} \leq p \leq n^{2-\zeta}$, where $\zeta \in (0, 1)$ can be chosen arbitrarily small. LB stands for Lower bound and UB stands for upper bound.

k_0	Δ	$\rho_{g, \gamma}^{*2}[k_0, \Delta]$
$k_0 \leq p^{1/2-\zeta}$	$1 \leq \Delta \leq p^{1/2-\zeta} \wedge k_0$	$\frac{\Delta \log(p)}{n}$
	$p^{1/2+\zeta} \wedge k_0 \leq \Delta \leq p - k_0$	$\frac{\sqrt{p}}{n}$
$p^{1/2+\zeta} \leq k_0 \leq c_\gamma \frac{n}{\log(p)}$	$1 \leq \Delta \leq k_0 p^{-\zeta}$	$\frac{\Delta \log(p)}{n}$
	$k_0 \leq \Delta \leq p - k_0$	LB : $\frac{k_0}{n \log(p)}$ UB : $\frac{k_0 \log(p)}{n}$

for the support of θ^* by relying on the non-convex penalized least-square estimator MCP [49]. Then, a test based on the restricted least-squares estimator applied to the selected subset is shown to achieve the desired separation distance. We also introduce an alternative test based on an iterative application of a projected version of the square-root Lasso.

1.4 Other related work

Two recent works [30, 52] have among other things consider general testing problems that encompass the sparsity testing problem. These two contributions assess the quality of their tests according to the l_∞ separation distance (instead of l_2 as we do here) to the null hypothesis, i.e. $d_\infty(\theta^*; \mathbb{B}_0[k_0]) = \inf_{\theta \in \mathbb{B}_0[k_0]} \|\theta^* - \theta\|_\infty$. In their setting, the covariance Σ of the covariates is unknown but its inverse Σ^{-1} is assumed to be sparse (each row of Σ^{-1} has at most $n/\log(p)$ non-zero entries) so that it can be reasonably well estimated. In that setting, the computationally feasible test in [52] has a small type II error probability when $k_0 \log(p)$ is much smaller than $n^{1/4}$ and when $d_\infty(\theta^*; \mathbb{B}_0[k_0]) \geq c\sigma n^{-1/4}$.

In [30], Javanmard and Lee use a test based on the debiased Lasso. It achieves a small type I error probability. Whenever $(k_0 + \Delta) \log(p)$ is much smaller than \sqrt{n} , and also $d_\infty(\theta^*; \mathbb{B}_0[k_0]) \geq c\sigma \sqrt{\log(p)/n}$, its type II error probability is also small. Translating these results in the l_2 separation distance setting, we observe that this test achieves a squared separation distance $\Delta \log(p)/n$ which, in view of Table 2, is optimal for small Δ . Their approach could be used instead of ours in their setting. However, we stress out that they achieve this bound to the price of considering a much more restricted class of covariance matrices than $\mathcal{U}(\eta)$ - they need that each row of Σ^{-1} is at most $n/\log(p)$ sparse, while $\mathcal{U}(\eta)$ contains all matrices Σ that have their spectrum contained in $[\eta^{-1}, \eta]$.

A recent line of work has focused on testing the nullity of a given subset of coordinates of θ^* (e.g. [6, 9, 32, 44, 51, 53, 54]), but both the settings and the methodology are quite different.

1.5 Notation

For any positive integer d and $u \in \mathbb{R}^d$, we write $\mathcal{S}(u) = \{i : u_i \neq 0\}$ for the support of a vector u . For $u \in \mathbb{R}^d$ and $S \subset [d]$, we write $u_S = (u_i \mathbf{1}_{i \in S})_i$ for the vector in \mathbb{R}^d whose values outside S have been set to 0. For a vector γ , $\gamma_{(i)}$ stands for its i -th largest (in absolute value) entry. Given $S \subset \{1, \dots, p\}$, \bar{S} stands for its complement.

In the sequel, c, c_1, c' denote numerical positive constants that may vary from line to line. Given some quantity δ , c_δ stands for a positive constant possibly depending on δ that may vary from line to line. Underlined constant such as $\underline{c}, \underline{c}^{(1)}$ do not vary in the paper.

Let $a, b \in \mathbb{R}$ be two functions that may depend on several quantities such as n, p, Δ, k_0 and let $u \in \mathbb{R}$. We write $a \lesssim_u b$ (resp. $a \approx_u b$) if there exists a constant $c_u > 0$ that depends only on u (resp. two constants $c_u^+, c_u^- > 0$ that depend only on u) such that $a \leq c_u b$ (resp. such that $c_u^- b \leq a \leq c_u^+ b$).

For $x > 0$, $\lfloor x \rfloor$ (resp. $\lceil x \rceil$) stands for the largest (resp. smallest) integer which is less (resp. greater) or equal to x . Also, \log_2 stands for the binary logarithm. Finally, $\bar{\Phi}$ stands for the tail distribution function of a standard normal distribution.

2 Independent setting

To simplify the notation, we denote $\mathbb{P}_{\theta^*, \sigma}$ the distribution of the data when Σ is the identity matrix. Recall that we are especially interested in the high-dimensional setting. This is why we shall sometimes assume that $p \geq n$ or even $p \geq n^{1+\zeta}$ for some $\zeta > 0$ arbitrarily small.

2.1 Minimax lower bound

As a starting point, we prove that, when the size k_0 of the null hypothesis is too large, consistent testing is impossible. Indeed, assume that $k_0 \geq n$. Then, for any $(Y, \mathbf{X}) \in \mathbb{R}^n \times \mathbb{R}^{n \times p}$ such that $\text{Rank}(\mathbf{X}) \geq n$, there exists $\theta \in \mathbb{B}_0[k_0]$ that perfectly fits this sample ($Y = \mathbf{X}\theta$) and it is therefore impossible to decipher whether θ^* is k_0 -sparse or not. The following proposition formalizes this observation.

Proposition 1. *If $k_0 \geq n$, then, for any $\gamma < 1/2$, and $1 \leq \Delta \leq p - k_0$, we have $\rho_\gamma^*[k_0, \Delta] = \infty$.*

In the sequel, we therefore restrict ourselves to the case where $k_0 < n$. The next theorem provides a lower bound for the minimax separation distance of the sparsity testing problem.

Theorem 1. *Assume that $p \geq 2n$. There exist positive numerical constants c_1 – c_5 such that the following holds for all $\gamma \leq 0.06$ and for all $p \geq c_1$. For $1 \leq \Delta \leq p - k_0$, one has*

$$\rho_\gamma^{*2}[k_0, \Delta] \geq c_1 \begin{cases} \min \left[\frac{1}{\sqrt{n}} + \frac{k_0}{n} \log \left[1 + \frac{\sqrt{p}}{k_0} \right], \frac{\Delta}{n} \log \left(1 + \frac{\sqrt{p}}{\Delta} \right) \right] & \text{if } 0 \leq k_0 \leq \sqrt{p} \wedge n ; \\ \frac{\Delta}{n} \frac{\log^2 \left[1 + \sqrt{\frac{k_0}{\Delta}} \right]}{\log(p)} & \text{if } \sqrt{p} < k_0 < n . \end{cases} \quad (6)$$

Furthermore, if $p \geq c_2 n^2$ and $k_0 \geq c_3 n / \log(\sqrt{p}/n)$, then

$$\rho_\gamma^{*2}[k_0, \Delta] \geq c_4 \frac{\Delta \wedge k_0}{n} \log \left(2 \vee \frac{\sqrt{p}}{k_0} \right) e^{c_5 \frac{k_0}{n} \log \left(1 + \frac{\sqrt{p}}{k_0} \right)}, \quad (7)$$

for all $1 \leq \Delta \leq p - k_0$.

In particular, (7) entails that the sparsity testing problem turns out to be extremely difficult in the regime $n/\log(p) \lesssim k_0 \lesssim n$ (at least when $p \geq n^2$).

The different regimes in (6) will be discussed together with the upper bounds at the end of the section. Let us shortly comment on the proof of Theorem 1. The functional $\rho_\gamma^*[k_0, \Delta]$ is (almost) nondecreasing with respect to k_0 . As a consequence, the lower bound $\frac{\Delta}{n} \log \left(1 + \frac{\sqrt{p}}{\Delta} \right)$ is a straightforward consequence of the analysis of the detection problem e.g. in [28].

The two lower bounds $\frac{1}{\sqrt{n}} + \frac{k_0}{n} \log \left[1 + \frac{\sqrt{p}}{k_0} \right]$ and (7) are based on a reduction argument. The proof stems from the fact it is impossible to decipher between two sets of hypothesis if these two sets of hypotheses are almost indistinguishable from a third party hypothesis. Here, the third party hypothesis corresponds to $\theta^* = 0$ and a tailored noise variance $\sigma' > \sigma$. Plugging minimax lower bounds for detection with *unknown* variance allows us to get the desired rate. See the proof for more details.

In fact, it is most challenging to prove the minimax lower bound in the regime $\Delta > k_0 > \sqrt{p}$ as we cannot apply any reduction technique to signal detection problem and we need to take into account that both the null and the alternative hypotheses are composite. As for the Gaussian sequence model [17], we use a general moment matching technique [38], but the non-orthogonal design matrix \mathbf{X} makes the computations more tricky.

2.2 Testing procedures

In this subsection, we fix α and $\delta \in (0, 1)$. We now introduce three testing procedures whose combination leads to matching the previous minimax lower bound.

Without loss of generality, we assume that n is divisible by 3 and we divide the sample (Y, \mathbf{X}) into three subsamples $(Y^{(1)}, \mathbf{X}^{(1)})$ and $(Y^{(2)}, \mathbf{X}^{(2)})$ and $(Y^{(3)}, \mathbf{X}^{(3)})$ of equal size $m = n/3$. For $i = 1, 2, 3$, we write $\mathbb{P}_{\theta^*, \sigma}^{(i)}$ for the probability according to the i -th sub-sample. In fact, some of the tests introduced below only use the first two subsamples. Nevertheless, we use three subsamples throughout the paper to simplify the presentation.

To characterize the performances of the testing procedures, we shall control the type I error probability uniformly over the null hypothesis and control the type II error probability on some 'large' parameter subset of the alternative. To simplify the statements of the results we shall refer to these two properties as **(P1)** and **(P2)** as defined below.

Property P1. A test ϕ satisfies **(P1)[α]** if its type I error probability is less than or equal to α , that is $\sup_{\theta^* \in \mathbb{B}_0[k_0]} \mathbb{P}_{\theta^*, \sigma}[\phi = 1] \leq \alpha$

Property P2. A test ϕ satisfies **(P2)[β]** on a set Θ if its type II error probability is uniformly less than or equal to β , uniformly on Θ , that is $\inf_{\theta^* \in \Theta} \mathbb{P}_{\theta^*, \sigma}[\phi = 1] \geq 1 - \beta$

Following the discussion in the previous subsection, we restrict our attention to sparsities k_0 that are less than $n/\log(p)$. This is formalized in the following condition **(A)[α]** where $\underline{c}^{(\mathbf{A})}, \underline{c}^{(\mathbf{A})'}$ are numerical constants (respectively small enough for $\underline{c}^{(\mathbf{A})}$ and large enough for $\underline{c}^{(\mathbf{A})'}$) whose values are constrained in Propositions 2–5.

$$\mathbf{(A)[\alpha]} \quad (k_0 \vee 1) \log\left(\frac{p}{\alpha}\right) + \log^2\left(\frac{p}{\alpha}\right) \leq \underline{c}^{(\mathbf{A})} n \quad \text{and} \quad p \geq \underline{c}^{(\mathbf{A})'}$$

2.2.1 Test $\phi^{(t)}$ based on a l_∞ estimation of θ^*

The first test aims at detecting whether θ^* contains at least $k_0 + 1$ 'large' entries. In order to do so, we need to build a reasonable l_∞ estimator of θ^* . Note that estimators based on the debiased Lasso have already been proved to achieve such a property (see e.g. [32]) in some settings. For the sake of completeness and as a gentle introduction to more challenging settings, we introduce here a slightly different estimator.

As a first step, we rely on a square-root Lasso [4] estimator based on the first subsample. From the design matrix $\mathbf{X}^{(1)}$, we build its column normalized modification $\mathbf{T}^{(1)}$ by

$$\mathbf{T}^{(1)} = \left(\mathbf{X}_{\cdot,1}^{(1)} / \|\mathbf{X}_{\cdot,1}^{(1)}\|_2, \dots, \mathbf{X}_{\cdot,p}^{(1)} / \|\mathbf{X}_{\cdot,p}^{(1)}\|_2 \right).$$

Set $\lambda = 2\sqrt{\Phi^{-1}(\delta/(4p))}$. The square-root Lasso estimator is then defined by

$$\widehat{\theta}_{SL,N} \in \arg \min \|Y^{(1)} - \mathbf{T}^{(1)}\theta\|_2 + \lambda\|\theta\|_1 ; \quad (\widehat{\theta}_{SL})_i = (\widehat{\theta}_{SL,N})_i / \|\mathbf{X}_{:,i}^{(1)}\|_2, \quad i = 1, \dots, p. \quad (8)$$

In this section, we could replace the square-root Lasso estimator by a classical Lasso estimator since the noise level σ is known. Also, the design is normalized for the purpose of simplifying some proof arguments, but the results remain valid (with slightly different constants) with the unnormalized design matrix $\mathbf{X}^{(1)}$.

Then, given $\widehat{\theta}_{SL}$, we use the second sample to improve the estimation of θ^* . The estimator $\widetilde{\theta}_{\mathbf{I}}$ is based on the empirical raw correlations between the covariates and the residuals.

$$\widetilde{\theta}_{\mathbf{I}} = \frac{1}{m} \mathbf{X}^{(2)T} (Y^{(2)} - \mathbf{X}^{(2)}\widehat{\theta}_{SL}) + \widehat{\theta}_{SL}. \quad (9)$$

Since the design is independent, $\widetilde{\theta}_{\mathbf{I}}$ is an unbiased estimator of θ^* . It is not hard to show (see the proof of the next proposition) that, under weak assumptions, this estimator satisfies $\|\widetilde{\theta}_{\mathbf{I}} - \theta^*\|_\infty \lesssim c\sigma\sqrt{\log(p)/n}$ with high probability. This is why we define the test $\phi^{(t)}$ rejecting the null if $|(\widetilde{\theta}_{\mathbf{I}})_{(k_0+1)}| \geq \underline{c}^{(t)}\sigma\sqrt{\log(p/\alpha)/n}$, where a suitable value for the constant $\underline{c}^{(t)}$ is defined in the proof of Proposition 2 below. This test is powerful when θ^* contains at least $k_0 + 1$ large entries. This is formalized in the following proposition.

Proposition 2. *There exist numerical constants $\underline{c}^{(t)}$, c and c' such that the following holds under Condition $(\mathbf{A}[\alpha \wedge \beta \wedge \delta])$. The test $\phi^{(t)}$ satisfies $(\mathbf{P1}[\alpha + \delta])$ and $(\mathbf{P2}[\beta + \delta])$ on the collections*

$$\mathbb{B}_0[k_0 + \Delta] \cap \left\{ \theta^*, \quad |\theta_{(k_0+1)}^*| \geq c\sigma\sqrt{\frac{1}{n} \log\left(\frac{p}{\alpha \wedge \beta}\right)} \right\}, \quad (10)$$

with $1 \leq \Delta \leq c'n/\log(p/\delta)$.

Again, we emphasize that similar performances are achieved by the debiased Lasso test of Javanmard and Lee [30].

2.2.2 Test $\phi^{(x)}$ based on the l_2 norm of the residuals

The second test is also simple. We heavily rely on the knowledge of the noise level σ . In the detection setting ($k_0 = 0$), [1, 28] consider a test rejecting the null when the squared norm $\|Y\|_2^2/(n\sigma^2)$ is large compared to one. Indeed, in expectation, $\|Y\|_2^2/(n\sigma^2)$ is equal to $\|\theta^*\|_2^2/\sigma^2 + 1$. Here, we have to adapt this statistic as $\|\theta^*\|_2^2$ is unknown under the null.

First, we project the square-root Lasso estimator $\widehat{\theta}_{SL}$ onto the parameter set corresponding to the null hypothesis. More precisely, we define $\widetilde{\theta}_{SL,k_0} = \arg \min_{\theta \in \mathbb{B}_0[k_0]} \|\widehat{\theta}_{SL} - \theta\|_2^2$. In other words, $\widetilde{\theta}_{SL,k_0}$ is obtained from $\widehat{\theta}_{SL}$ by thresholding its $(p - k_0)$ smallest entries to zero. Then, given $\widetilde{\theta}_{SL,k_0}$, we use the second sample to assess whether θ^* is significantly different from $\widetilde{\theta}_{SL,k_0}$. Define the residuals vectors $\widehat{R}_{k_0} = Y^{(2)} - \mathbf{X}^{(2)}\widetilde{\theta}_{SL,k_0}$ and, for $R \in \mathbb{R}^m$, the statistic $Z_\chi[R] = \frac{\|R\|_2^2}{m\sigma^2} - 1$.

Take the threshold $v_{\alpha,\delta}^{(x)} = \sqrt{\frac{\log(1/\alpha)}{m} + \frac{(k_0 \vee 1)\log(p/\delta)}{m}}$, we consider the test $\phi^{(x)}$ rejecting the null hypothesis when $Z_\chi[\widehat{R}_{k_0}] > \underline{c}^{(x)}v_{\alpha,\delta}^{(x)}$, where the numerical constant $\underline{c}^{(x)}$ is introduced in the proof of the following proposition.

Proposition 3. *There exist numerical constants $\underline{c}^{(x)}$ and c and such that the following holds under Condition $(\mathbf{A}[\alpha \wedge \beta \wedge \delta])$. The test $\phi^{(x)}$ satisfies $(\mathbf{P1}[\alpha + \delta])$ and $(\mathbf{P2}[\beta])$ on the collection*

$$\left\{ \theta^*, \quad d_2^2[\theta^*; \mathbb{B}_0[k_0]] \geq c\sigma^2 \left[\frac{k_0 \vee 1}{n} \log(p/\delta) + \sqrt{\frac{\log(2/(\alpha \wedge \beta))}{n}} \right] \right\}. \quad (11)$$

It turns out that a combination of $\phi^{(t)}$ and $\phi^{(x)}$ is matching the minimax lower bound of Theorem 1 when $k_0 \leq \sqrt{p}$. For larger null hypotheses, we need to rely on more intricate tests that are discussed in the next section.

2.2.3 Test $\phi^{(f)}$ based on the empirical Fourier transform of the raw covariances

In the Gaussian sequence framework ($p = n$ and $\mathbf{X} = \mathbf{I}_p$), [17] have recovered the optimal separation distance using test based on the empirical Fourier transform of the data. In this section, we adapt this approach in the linear regression model.

Conditionally to Y , it is shown in the proof of Proposition 4 below that the normalized raw covariances $\mathbf{X}^T Y / \|Y\|_2$ follow a normal distribution with mean $\theta^* \|Y\|_2 / [\sigma^2 + \|\theta^*\|_2^2]$ and variance $\mathbf{I}_p - \theta^* \theta^{*T} / [\sigma^2 + \|\theta^*\|_2^2]$. Since $\|Y\|_2^2$ is concentrated around $n[\sigma^2 + \|\theta^*\|_2^2]$ and assuming that $\|\theta^*\|_2^2$ is small compared to σ^2 , this implies that the raw covariances are almost distributed as a normal distribution with mean $\sqrt{n}\theta^*/\sigma$ and covariance \mathbf{I}_p . This observation leads us to adapt the Fourier tests of [17] in our setting through raw covariances.

The purpose of the empirical Fourier transform statistic considered in [17] (but see also [33, 34] for previous work), is to approximate the discontinuous function $\sum_{i=1}^p \mathbf{1}_{\theta_i^* \neq 0}$. First, introduce, for $s > 0$, the function

$$\varphi(s; x) = \int_{-1}^{+1} (1 - |\xi|) \cos(\xi s x) e^{\xi^2 s^2 / 2} d\xi. \quad (12)$$

For $Z \sim \mathcal{N}(a, 1)$, standard computations lead to $\mathbb{E}[\varphi(s; Z)] = 2 \frac{1 - \cos(sa)}{(sa)^2} =: g(sa)$. In particular, the function g takes values in $[0, 1]$ with $g(0) = 0$ and $\lim_{|a| \rightarrow \infty} g(sa) = 1$ (see [17]). In some way, $g(sa)$ is a smooth approximation to $\mathbf{1}_{a \neq 0}$. The larger s is, the closer $g(sa)$ is to the indicator function. However, $\varphi(s; Z)$ exhibit a higher variance for large s .

The conditional distribution of $\mathbf{X}^T Y / \|Y\|_2$ is close to a normal distribution with mean $\sqrt{n}\theta^*/\sigma$ and variance-covariance matrix \mathbf{I}_p , provided that $\|\theta^*\|_2^2$ is small compared to σ^2 . Hence, it would be tempting to use a statistic of the form $\sum_{i=1}^p \varphi(s; (\mathbf{X}^T Y)_i / \|Y\|_2)$, which in expectation would be close to $\sum_{i=1}^p g(s\theta_i^*)$, which in turn would approximate $\|\theta^*\|_0$. Unfortunately, large coordinates $|\theta_i^*|$ may perturb the concentration of the statistic since the true conditional covariance of $\mathbf{X}^T Y / \|Y\|_2$ is $\mathbf{I}_p - \theta^* \theta^{*T} / [\sigma^2 + \|\theta^*\|_2^2]$. To address this technical issue, we first correct θ^* by removing its large coefficients.

As in Subsection 2.2.1, the first two samples are respectively dedicated to building the Lasso estimator $\hat{\theta}_{SL}$ and the debiased estimator $\tilde{\theta}_{\mathbf{I}}$. If $|\tilde{\theta}_{\mathbf{I}}|_{(k_0+1)} > \underline{c}^{(t)} \sigma \sqrt{\log(2p/\alpha)/n}$, then the test $\phi^{(f)}$ introduced below rejects the null hypothesis, otherwise we define $\bar{\theta}_{\mathbf{I}}$ as

$$\bar{\theta}_{\mathbf{I},i} = \tilde{\theta}_{\mathbf{I},i} \mathbf{1}_{\{|\tilde{\theta}_{\mathbf{I},i}| > \underline{c}^{(t)} \sigma \sqrt{\frac{\log(2p/\alpha)}{n}}\}}. \quad (13)$$

In Subsection 2.2.1, we argued that, with high probability, $\|\tilde{\theta}_{\mathbf{I}} - \theta^*\|_\infty \leq \underline{c}^{(t)} \sigma \sqrt{\log(2p/\alpha)/n}$. As a consequence, $\|\bar{\theta}_{\mathbf{I}} - \theta^*\|_\infty \leq 2\underline{c}^{(t)} \sigma \sqrt{\frac{\log(2p/\alpha)}{n}}$ and the support of $\bar{\theta}_{\mathbf{I}}$ is included in that of θ^*

Finally, we use the third subsample to compute the corrected raw covariances $W_j = \mathbf{X}_j^{(3)T} \bar{\mathbf{Y}}^{(3)}$ with $\bar{\mathbf{Y}}^{(3)} = Y^{(3)} - \mathbf{X}^{(3)} \bar{\theta}_{\mathbf{I}}$ relative to the linear regression model with parameter $\theta^* - \bar{\theta}_{\mathbf{I}}$. Then, following the above heuristic explanation, we consider the statistic

$$Z_f := \sum_{j=1}^p \mathbf{1}_{(\bar{\theta}_{\mathbf{I}})_j = 0} \varphi(s; \frac{W_j}{\|\bar{\mathbf{Y}}^{(3)}\|_2}) + \mathbf{1}_{(\bar{\theta}_{\mathbf{I}})_j \neq 0},$$

with tuning parameter $s = \sqrt{\log(e \frac{k_0}{\sqrt{p}})} \vee 1$. For j in the support of $\bar{\theta}_I$, we are already confident that θ_j^* is non zero and we do not have to rely on φ . Finally, the test $\phi^{(f)}$ rejects the null when $Z_f \geq k_0 + v_\alpha^{(f)}$ with $v_\alpha^{(f)} = s^2/5 + se^{s^2/2} \sqrt{2p \log(2/\alpha)}$.

In comparison to the original statistic of [17] for the Gaussian sequence model, we use here a slightly smaller tuning parameter s and the threshold $v_\alpha^{(f)}$ has an additional term $s^2/5$.

Proposition 4. *There exist constants c, c_α, c'_α , and c''_α such that the following holds under Condition $(\mathbf{A}[\alpha \wedge \delta])$. The test $\phi^{(f)}$ satisfies $(\mathbf{P1}[\alpha + \delta])$ and $(\mathbf{P2}[\alpha + \delta + e^{-n/27}])$ on the collection of parameters θ^* satisfying $\|\theta^*\|_0 \leq cn/\log(p/\delta)$, $d_2^2[\theta^*; \mathbb{B}_0[k_0]] \leq \sigma^2$ and at least one of the two following conditions.*

$$|\theta_{(k_0+q)}^*| \geq c_\alpha \sigma \sqrt{\frac{k_0}{qn \log(1 + k_0/\sqrt{p})}}, \text{ for some } q \geq c'_\alpha \left((k_0^{3/4} p^{1/8}) \vee \sqrt{p} \right); \quad (14)$$

$$\sum_{i=1}^p \left[\theta_i^{*2} \wedge \frac{\sigma^2}{n \log(1 + k_0/\sqrt{p})} \right] \geq c''_\alpha \sigma^2 \frac{k_0}{n \log(1 + k_0/\sqrt{p})}. \quad (15)$$

The test $\phi^{(f)}$ rejects the null hypothesis when there are many small non-zero coefficients in θ^* . In particular, if θ^* contains $2k_0 > 2\sqrt{p}$ coefficients of order $\sigma(n \log(p))^{-1/2}$, then the null hypothesis is rejected with high probability. Note that $\sigma(n \log(p))^{-1/2}$ is much smaller than the value needed to recover the position of these non-zero coefficients, which is of the order $\sigma\sqrt{\log(p)/n}$. This behavior is reminiscent of the minimax lower bound in Theorem 1, where the squared separation distance is proven to be at least of the order $k_0/[n \log(p)]$ for $\Delta \geq k_0 \geq \sqrt{p}$.

When there are a few entries in θ^* that are neither large nor small - see below for more precisions, it turns out that $\phi^{(f)}$ only matches the minimax lower bound up to some $\log \log(p)$ multiplicative factor. To address this issue we need to introduce an additional test $\phi^{(i)}$.

2.2.4 Intermediary regime: Test $\phi^{(i)}$ based on the empirical Fourier transform of the raw covariance

In this subsection, we focus on entries θ_i^* that are neither large (with respect to $\sigma\sqrt{\log(p)/n}$) as in the analysis $\phi^{(t)}$ nor small (with respect to $\sigma\sqrt{1/(n \log(p))}$) as in the analysis of $\phi^{(f)}$. This setting turns out to be relevant for large k_0 only and we assume henceforth that $k_0 \geq 2^{11}\sqrt{p}$. As in the previous section, we adapt a test from [17] in the Gaussian sequence setting by applying the empirical Fourier transform to the raw covariances.

Given two tuning parameters r and l , define the function

$$\eta_{r,w}(x) = \frac{r}{(1 - 2\bar{\Phi}(r))} \int_{-1}^1 \frac{e^{-r^2 \xi^2/2}}{\sqrt{2\pi}} e^{\xi^2 w^2/2} \cos(\xi w x) d\xi. \quad (16)$$

and the statistic

$$V(r, w) = \sum_{j=1}^p \mathbf{1}_{(\bar{\theta}_I)_j=0} [1 - \eta_{r,w}(W_j / \|\bar{Y}^{(3)}\|_2)] + \mathbf{1}_{(\bar{\theta}_I)_j \neq 0}.$$

In order to get a grasp of this statistic let us consider the expectation of $\eta_{r,w}(X)$ for $X \sim \mathcal{N}(x, 1)$. Simple computations (see [17]) lead to $\mathbb{E}[1 - \eta_{r,w}(X)] = 1 - \frac{1}{1 - 2\bar{\Phi}(r)} \int_{-r}^r \phi(\xi) \cos(\xi x \frac{w}{r}) d\xi$, which for large r , is close to $1 - \exp(-x^2 \frac{w^2}{2r^2})$. Thus, in contrast to the population function g introduced in

the previous subsection, which converges to 1 at a quadratic rate, this function converges to one at an exponential rate, thereby better handling moderate values of θ_i^* . The downside of using this statistic is that $\mathbb{E}[1 - \eta_{r,w}(X)]$ does not lie in $[0, 1]$.

The test $\phi^{(i)}$ is an aggregation of multiple tests based on the statistics $V(r, w)$ for different tuning parameters r and w . Define $l_0 = \lceil k_0^{4/5} p^{1/10} \rceil$ and the dyadic collection $\mathcal{L}_0 = \{l_0, 2l_0, 4l_0, \dots, l_{\max}\}$ where $l_{\max} = 2^{\lceil \log_2(k_0/l_0) \rceil} l_0/4 \leq k_0/4$. Note that \mathcal{L}_0 is not empty if $k_0 \geq 2^{11} \sqrt{p}$ and p is large enough. Given any $l \in \mathcal{L}_0$, define

$$r_l = \sqrt{2 \log\left(\frac{k_0}{l}\right)}; \quad w_l = \sqrt{\log\left(\frac{l}{\sqrt{p}}\right)}. \quad (17)$$

Then, the test $\phi^{(i)}$ rejects the null hypothesis if, for some $l \in \mathcal{L}_0$,

$$V(r_l, w_l) \geq k_0 + l + v_{\alpha, l}^i \quad \text{where} \quad v_{\alpha, l}^i = \frac{e^{1/2}}{2} \omega_l^2 + \sqrt{2ln^{1/2} \log\left(\frac{\pi^2[1 + \log_2(l/l_0)]^2}{6\alpha}\right)}. \quad (18)$$

In comparison to the test in [17], the collection of tuning parameters \mathcal{L}_0 is slightly narrower and the threshold $v_{\alpha, l}^i$ has an additional corrective term of the order of ω_l^2 .

Proposition 5. *There exist positive constants c, c_α, c'_α such that the following holds under Condition $(\mathbf{A}[\alpha \wedge \delta])$. The test $\phi^{(i)}$ satisfies $(\mathbf{P1}[\alpha + \delta])$ and $(\mathbf{P2}[\alpha + \delta + e^{-n/27}])$ on the collection of parameters θ^* satisfying $\|\theta^*\|_0 \leq cn/\log(p/\delta)$, $d_2^2[\theta^*; \mathbb{B}_0[k_0]] \leq \sigma^2$ and*

$$|\theta_{(k_0+q)}^*| \geq c_\alpha \sigma \frac{1 + \log\left(\frac{k_0}{q \wedge k_0}\right)}{\sqrt{n \log\left(1 + \frac{k_0}{\sqrt{p}}\right)}}, \quad \text{for some } q \geq c'_\alpha k_0^{4/5} p^{1/10}. \quad (19)$$

In Comparison to Condition (14) for Proposition 4, $|\theta_{(k_0+q)}^*|$ is possibly much smaller than for $\phi^{(f)}$ in the regime where $k_0^{4/5} p^{1/10} \lesssim_\alpha q \lesssim_\alpha \sqrt{p}$.

2.2.5 Aggregated test

To conclude this section, we evaluate the performances of the combination of all the previous tests. In fact, $\phi^{(i)}$ is only defined in the large k_0 regime. We take the convention that $\phi^{(i)}$ is a trivial test that always accepts the null hypothesis in the small k_0 regime. Consider the aggregated test

$$\phi^{(ag)} = \max(\phi^{(t)}, \phi^{(x)}, \phi^{(f)}, \phi^{(i)}, \mathbf{1}\{d_2^2(\widehat{\theta}_{SL}, \mathbb{B}_0[k_0]) \geq \sigma^2/2\}).$$

The last test $\mathbf{1}\{d_2^2(\widehat{\theta}_{SL}, \mathbb{B}_0[k_0]) \geq \sigma^2/2\}$ is introduced for technical purpose to handle very dense alternatives ($\|\theta^*\|_0 \geq cn/\log(p/\delta)$).

Theorem 2. *Let $\delta \in (0, 1)$ and $\varsigma \in (0, 1)$. There exists positive constants c_ς and $c_{\varsigma, \delta}$ such that the following holds. Assume that $p \geq c_\varsigma$ and that Condition $(\mathbf{A}[\delta \wedge \delta])$ is satisfied. Define*

$$\rho_{k_0, \Delta, \varsigma}^2 = \begin{cases} \min\left[\frac{\Delta}{n} \log(p), \frac{1}{\sqrt{n}} + \frac{k_0}{n} \log(p)\right] & \text{if } 0 \leq k_0 \leq p^{1/2-\varsigma}; \\ \min\left[\frac{\Delta \log(p)}{n}, \frac{k_0}{n \log(p)}\right] & \text{if } k_0 > p^{1/2+\varsigma}. \end{cases} \quad (20)$$

The test $\phi^{(ag)}$ satisfies $(\mathbf{P1}[\delta + 4\alpha])$ and $(\mathbf{P2}[\delta + \alpha + e^{-n/27}])$ on the collection of parameters

$$\mathbb{B}_0[k_0 + \Delta] \cap \{\theta^*, d_2^2[\theta^*; \mathbb{B}_0[k_0]] \geq c_{\delta, \varsigma} \rho_{k_0, \Delta, \varsigma}^2\}$$

with $1 \leq \Delta \leq p - k_0$.

The case $k_0 + \Delta \leq cn/\log(p/\delta)$ is a simple corollary of the previous results, whereas the dense case $k_0 + \Delta > cn/\log(p/\delta)$ requires further work.

To further compare this result with the minimax lower bound of Theorem 1, we assume that $p \geq n^{1+\zeta}$ for some $\zeta > 0$. Recall that we also suppose $k_0 \leq cn/\log(p)$. From Theorems 1 and 2, we deduce that

Case 1: $k_0 \leq p^{1/2-\kappa}$ with an arbitrary $\kappa \in (0, 1/2)$.

$$\rho_\gamma^{*2}[k_0, \Delta] \asymp_{\gamma, \kappa, \varsigma} \begin{cases} \frac{\Delta}{n} \log(p) & \text{if } \Delta \leq \frac{\sqrt{n}}{\log(p)} + k_0 ; \\ \frac{1}{\sqrt{n}} + \frac{k_0 \log(p)}{n} & \text{if } \Delta > \frac{\sqrt{n}}{\log(p)} + k_0 . \end{cases}$$

Case 2: $k_0 \geq p^{1/2+\kappa}$ with an arbitrary $\kappa \in (0, 1/2)$. For any $\varsigma \in (0, 1/2)$ arbitrarily small, we have

$$\rho_\gamma^{*2}[k_0, \Delta] \asymp_{\gamma, \kappa, \zeta, \varsigma} \begin{cases} \frac{\Delta}{n} \log(p) & \text{if } \Delta < k_0 p^{-\varsigma} ; \\ \frac{k_0}{n \log(p)} & \text{if } \Delta \geq k_0 , \end{cases}$$

and that all these bounds are simultaneously achieved by the test $\phi^{(ag)}$. As a consequence, $\phi^{(ag)}$ is simultaneous minimax over all k_0 and all Δ except in the regimes when k_0 is close to \sqrt{p} or when Δ is close to k_0 , in which case, there is possibly a polylogarithmic difference between the minimax lower and upper bounds.

Proof. If $k_0 \leq p^{1/2-\kappa}$ and $p \geq n^{1+\zeta}$, then $\log(1 + \sqrt{p}/k_0) \asymp_\kappa \log(p)$. Hence (6) in Theorem 1 ensures that the square minimax separation distance is at least of the order of $\min(\frac{\Delta}{n} \log(1 + \frac{\sqrt{p}}{\Delta}), \frac{1}{\sqrt{n}} + \frac{k_0 \log(p)}{n})$. The first term is (up to numerical constants) larger than second one when $\Delta \geq k_0 \vee \sqrt{n}$. For $\Delta \leq k_0 \vee \sqrt{n}$, we have $\log(\sqrt{p}/\Delta) \asymp_{\kappa, \zeta} \log(p)$. Hence, the square minimax separation distance is at least of the order of $\min(\frac{\Delta}{n} \log(p), \frac{1}{\sqrt{n}} + \frac{k_0 \log(p)}{n})$ which matches the upper bound of Theorem 2. If $k_0 \geq p^{1/2+\kappa}$, $p \geq n^{1+\zeta}$, and $\Delta \leq k_0 p^{-\varsigma}$, then $\log(1 + \sqrt{k_0/\Delta}) \asymp_{\kappa, \zeta, \varsigma} \log(p)$ and Theorem 6 ensures that the square minimax separation distance is at least of the order of $\Delta \log(p)/n$, matching again Theorem 2. When $\Delta \geq k_0$, $\Delta \log^2(1 + \sqrt{k_0/\Delta}) \geq ck_0$, and we deduce from Theorems 1 and 2 that the square minimax separation distance is of order of $k_0/[n \log(p)]$. \square

Let us summarize the different regimes

- If Δ is small - first result in Cases 1 and 2 - then the squared minimax separation distance ($\Delta \log(p)/n$) is the same as for signal detection ($k_0 = 0$). The upper bound can be achieved using any $\sqrt{\log(p)/n}$ l_∞ -consistent estimator of θ^* and simply counting the number of its large entries. In the independent setting such estimator is easily built using the raw correlation ($\tilde{\theta}_1$) between the variables and the response. Alternatively, one could use the debiased Lasso [31, 32, 44, 51] which is valid for a wider class of Σ .
- If Δ is large and k_0 is small - second result in Case 1 - then the squared minimax separation distance can be understood as the sum of the quantity $n^{-1/2}$ arising in signal detection and the complexity $k_0 \log(p)/n$ of the null hypothesis. The matching upper bound is achieved by computing the l_2 norm of the residuals when plugging a suitable estimator of θ^* . The upper bound was already obtained in [39] (for a computationally inefficient method) but the matching minimax lower bound is new.

- Finally, if Δ is large and k_0 is large - second result in Case 2 - then the minimax separation distance is highly non standard and depends on the complexity of the null hypothesis. Both the lower and upper bound are new. In some way, they both draw inspiration from the analysis [17] of the same problem in the Gaussian sequence framework.

In this paper, we focused on recovering the minimax separation distance in the the high-dimensional setting, namely we require $p \geq n^{1+\zeta}$, where $\zeta > 0$ is an arbitrarily small universal constant. Aside from this restriction, there are two gaps in our analysis:

- First, when $p \geq c_2 n^2$ and $k_0 \in (n/\log(p), n)$, our minimax lower bounds in Theorem 1 imply that the testing problem is almost impossible. However, for $p \leq c_2 n^2$, we did not manage to prove similar lower bounds. We conjecture that, for $p \leq c_2 n^2$ and $k_0 \in (n/\log(p), n)$, $\rho_\gamma^*[k_0, p]$ is huge, but we did not manage to prove it.
- Some poly-log terms mismatch between the upper and lower bounds arise when k_0 is close to \sqrt{p} - e.g. $k_0 \in [\sqrt{p} \log^{-\zeta}(p), \sqrt{p} \log^\zeta(p)]$ for some $\zeta > 0$ and when Δ gets close to k_0 from below - i.e. $k_0 p^{-\zeta} \leq \Delta \leq k_0$ for some arbitrarily small universal constant $\zeta > 0$. In that regime, we could improve our upper bounds by adapting some higher-criticism [23] procedures as it was done in the sequence model [17]. However, even with this new procedure this would not completely close the gap. We conjecture that our minimax bound (6) is not completely sharp in that regime (see its proof for a tentative explanation).

3 General Setting

In this section, we focus on the general setting where Σ is unknown and is only assumed to belong to some class $\mathcal{U}(\eta)$ (4) for some $\eta > 1$. The noise variance σ^2 is also assumed to be unknown.

3.1 Minimax lower bound

Obviously, $\rho_{g,\gamma}^*[k_0, \Delta]$ is at least as large as $\rho_\gamma^*[k_0, \Delta]$ since the covariance matrix Σ is unknown and \mathbf{I}_p belongs to $\mathcal{U}[\eta]$. Therefore, Theorem 1 in the previous section provides a lower bound on $\rho_{g,\gamma}^*[k_0, \Delta]$. It turns out that that this lower bound is sometimes loose and that the general setting is actually more challenging in some regimes as shown by the following proposition.

Proposition 6. *Assume that $p \geq 2n$. There exist positive numerical constants $c > 0$ and $c' > 0$ such that for $p \geq c_3$ and for all $\gamma \leq 0.06$, one has*

$$\rho_{g,\gamma}^{*2}[k_0, \Delta] \geq c \begin{cases} \frac{\Delta}{n} \log \left(1 + \frac{\sqrt{p}}{\Delta} \right) \exp \left[c' \frac{\Delta}{n} \log \left(1 + \frac{\sqrt{p}}{\Delta} \right) \right] & \text{if } 0 \leq k_0 \leq \sqrt{p} \wedge n ; \\ \frac{\Delta}{n} \frac{\log^2 \left[1 + \sqrt{\frac{k_0}{\Delta}} \right]}{\log(p)} & \text{if } \sqrt{p} \leq k_0 \leq n , \end{cases} \quad (21)$$

with $1 \leq \Delta \leq p - k_0$.

In fact, this result is a combination of Theorem 1 together with known minimax lower bounds for the detection problem ($k_0 = 0$) with unknown variance [45, 46].

In comparison to the independent setting, one cannot achieve anymore the rate $1/\sqrt{n} + k_0 \log(1 + \sqrt{p}/k_0)/n$. Most importantly, the testing problem becomes almost impossible for dense alternative ($\Delta \gtrsim n/\log(p)$) in the high-dimensional regime $p \geq n^2$.

3.2 Testing procedures

We cannot rely anymore on the test $\phi^{(x)}$ as the noise level is unknown nor on $\phi^{(t)}$ and $\phi^{(f)}$ as their reconstruction relies on the independence of the covariates.

As in the previous sections we introduce two properties (**gP1**) and (**gP2**) characterizing the type I and II error probabilities in this setting where the noise level σ and the covariance matrix Σ are unknown.

Property gP1. A test ϕ satisfies (**gP1** $[\alpha]$) if its type I error probability is less or equal to α , that is $\sup_{\theta^* \in \mathbb{B}_0[k_0]} \sup_{\sigma > 0} \sup_{\Sigma \in \mathcal{U}[\eta]} \mathbb{P}_{\theta^*, \sigma, \Sigma}[\phi = 1] \leq \alpha$.

Property gP2. A test ϕ satisfies (**gP2** $[\beta]$) on the collection Θ of parameters if its type II error probability is uniformly less or equal to β that is $\inf_{\sigma > 0, \Sigma \in \mathcal{U}[\eta]} \inf_{\theta^* \in \Theta} \mathbb{P}_{\sigma \theta^*, \sigma, \Sigma}[\phi = 1] \geq 1 - \beta$.

Note that that in the above bound θ^* is rescaled by σ for homogeneity purpose. As in Section 2, we restrict our attention to sparsities k_0 that are less than $n/\log(p)$. The numerical constants $\underline{c}^{(\mathbf{B})}$ and $\underline{c}^{(\mathbf{B})'}$ in the following condition are introduced in the proof of Proposition 7 and Corollary 1.

$$([\mathbf{B}[\alpha]) \quad (k_0 \vee 1)[1 + \log(p/\alpha)] + \log^3\left(\frac{1}{\alpha}\right) + \log(p) \log\left(\frac{1}{\alpha}\right) \leq \underline{c}_\eta^{(\mathbf{B})} n \quad \text{and} \quad p \geq \underline{c}^{(\mathbf{B})'}.$$

In this section, we divide the sample in two subsamples $(Y^{(0)}, \mathbf{X}^{(0)})$ and $(Y^{(1)}, \mathbf{X}^{(1)})$ of equal size $m = n/2$. As previously, we shall combine several tests to match the minimax lower bounds.

3.2.1 Test $\phi^{(u)}$ based on a U -statistic.

The first test is specific to the moderate regime $p \leq n^2$. For known σ , we introduced in the previous section a statistic relying on the observation that $\|Y\|_2^2/n - \sigma^2$ estimates well $\|\theta^*\|_2^2$. Then, relying on a good k_0 -sparse estimator $\tilde{\theta}_{SL, k_0}$ of θ^* and computing the square norm of the residuals, we estimate $\|\theta^* - \tilde{\theta}_{SL, k_0}\|_2^2$, which under the null, should be small. Here, we follow the same strategy by considering an estimator of the signal strength, still valid for unknown σ .

In [22], Dicker tackled the problem of estimating the signal strength $\|\theta^*\|_2^2$ in the setting where $\Sigma = \mathbf{I}_p$ and σ^2 is unknown. This led him to introduce the U -statistic $N = \frac{1}{n^2}[Y^T[\mathbf{X}\mathbf{X}^T - \frac{1}{n}\text{tr}[\mathbf{X}\mathbf{X}^T]\mathbf{I}_n]Y]$, which is unbiased and \sqrt{p}/n consistent. For general Σ , this statistic has later been shown to be concentrated around the quadratic form $\theta^{*T}\Sigma^2\theta^*$ (see [47, Sect.2.1]). As a consequence, one can rely on it to test the nullity of θ^* .

For composite null hypotheses, we use $(Y^{(1)}, \mathbf{X}^{(1)})$ to build $\tilde{\theta}_{SL, k_0}$ as in Subsection 2.2.2 and then compute the residuals \hat{R}_{SL} with respect to the the second sample, $\hat{R}_{SL} = Y^{(0)} - \mathbf{X}^{(0)}\tilde{\theta}_{SL, k_0}$. Finally, we define the normalized U -statistic $Z^{(u)}$ by

$$Z^{(u)} = \frac{\hat{R}_{SL}^T \left[\mathbf{X}^{(0)} \mathbf{X}^{(0)T} - \frac{1}{m} \text{tr}[\mathbf{X}^{(0)} \mathbf{X}^{(0)T}] \mathbf{I}_m \right] \hat{R}_{SL}}{\|\hat{R}_{SL}\|_2^2 (m+1)}. \quad (22)$$

Conditionally to $\tilde{\theta}_{SL, k_0}$, \hat{R}_{SL} is the response of a linear regression model with parameter $(\theta^* - \tilde{\theta}_{SL, k_0})$, variance $\sigma^2 \mathbf{I}_m$, and random design $\mathcal{N}(0, \Sigma)$. Hence, the second moment of each entry of \hat{R}_{SL} equals $\sigma^2 + (\theta^* - \tilde{\theta}_{SL, k_0})^T \Sigma (\theta^* - \tilde{\theta}_{SL, k_0})$ and $\|\hat{R}_{SL}\|_2^2/m$ is therefore close to $\sigma^2 + (\theta^* - \tilde{\theta}_{SL, k_0})^T \Sigma (\theta^* - \tilde{\theta}_{SL, k_0})$. Intuitively, the statistic $Z^{(u)}$ is therefore expected to be close to

$$\frac{(\theta^* - \tilde{\theta}_{SL, k_0})^T \Sigma^2 (\theta^* - \tilde{\theta}_{SL, k_0})}{\sigma^2 + (\theta^* - \tilde{\theta}_{SL, k_0})^T \Sigma (\theta^* - \tilde{\theta}_{SL, k_0})},$$

so that a large value for $Z^{(u)}$ suggests that θ^* is significantly different from a k_0 sparse vector. Setting the threshold

$$v_\alpha^{(u)} = \frac{(k_0 \vee 1) \log(p/\delta)}{m} + \frac{\sqrt{p \log(2/\alpha)}}{m}, \quad (23)$$

we consider the test $\phi^{(u)}$ rejecting the null hypothesis when $Z^{(u)} > \underline{c}_\eta^{(u)} v_\alpha^{(u)}$.

Proposition 7. *There exist three constants $\underline{c}_\eta^{(u)}$, c_η and c'_η such that the following holds under Condition $(\mathbf{B}(\alpha \wedge \beta \wedge \delta))$ and if $2n \leq p \leq c_\eta n^2 \log^{-1}(\frac{2}{\alpha \wedge \beta})$. The test $\phi^{(u)}$ satisfies $(\mathbf{gP1}[\alpha + \delta])$ and $(\mathbf{gP2}[\beta])$ on the collection*

$$\left\{ \theta^*, \quad d_2^2[\theta^*; \mathbb{B}_0(k_0)] \geq c'_\eta \left[\frac{\sqrt{p \log(2/(\alpha \wedge \beta))}}{n} + \frac{(k_0 \vee 1)}{n} \log\left(\frac{p}{\delta}\right) \right] \right\}. \quad (24)$$

3.2.2 Recovering the $\Delta \log(p)/n$ rate with variable selection

To achieve the $\Delta \log(p)/n$ rate, it would suffice to estimate θ^* at the l_∞ rate $\sigma \sqrt{\log(p)/n}$ as we did for the test $\phi^{(t)}$ in the previous section. However, we are unaware of any estimator achieving this rate uniformly over the class $\mathcal{U}(\eta)$ of covariance matrices Σ . For $k_0 \geq \sqrt{n}$, it is even proved that no such estimator exists [9].

Here, we adopt another strategy. We shall first estimate the support $\mathcal{S}(\theta^*)$ of θ^* and count the number of large entries of the least-squares estimator of θ^* restricted to the estimated support \widehat{S} . Of course, if $\widehat{S} = \mathcal{S}(\theta^*)$ with high probability, then the restricted least-squares estimator $\widehat{\theta}_{\widehat{S}}$ (see below for a definition) will be close to θ^* in l_∞ norm. Unfortunately, it is impossible for an estimator \widehat{S} to estimate exactly the support $\mathcal{S}(\theta^*)$, especially when θ^* contains arbitrarily small coordinates.

This is why we shall require that the estimator \widehat{S} satisfies a weaker property. Given $a > 0$, let $M(\mathbf{a}_1, \frac{\theta^*}{\sigma}) = |\{i, 0 < \frac{|\theta_i^*|}{\sigma} \leq \mathbf{a}_1 \sqrt{\log(p)/m}\}|$ be the number of small but non zero coefficients of θ^* . Below $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ refer to three positive quantities. Recall that \overline{S} is the complement of S .

Property $(\mathbf{S}[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3])$. A (possibly random) set S is said to satisfy this property if

$$|S| \leq \mathbf{a}_2 \|\theta^*\|_0; \quad \|\theta_{\overline{S}}^*\|_2^2 \leq \mathbf{a}_3^2 \sigma^2 M(\mathbf{a}_1, \frac{\theta^*}{\sigma}) \frac{\log(p)}{m}. \quad (25)$$

In other words, the cardinal of S is not too large compared to the sparsity of θ^* and the square norm of θ^* outside S is at most as large as that of the small entries of θ^* . Observe that the large entries of θ^* are not required to belong to S .

Then, given a set S , we consider the restricted least-square estimator and the plug-in variance estimators

$$\widehat{\theta}_{l_s, S} = \arg \min_{\theta : \mathcal{S}(\theta) \subset S} \|Y^{(0)} - \mathbf{X}^{(0)} \theta\|_2^2; \quad \widehat{\sigma}_S^2 = \frac{1}{m} \|Y^{(0)} - \mathbf{X}^{(0)} \widehat{\theta}_{l_s, S}\|_2^2. \quad (26)$$

For a vector $u \in \mathbb{R}^p$ and $c > 0$, $N[c; u] = |\{i : |u_i| \geq c \sqrt{\log(p)/m}\}|$ is the number of entries of u larger or equal (in absolute value) than $c \sqrt{\log(p)/m}$. Then, we define the test $\phi^{(th)}[S; c]$ rejecting the null if and only if $N[\underline{c}; \widehat{\theta}_{l_s, S} / \widehat{\sigma}_S] \geq k_0 + 1$, which means that $\widehat{\theta}_{l_s, S}$ contains at least $k_0 + 1$ large entries.

Theorem 3. *There exist constants c and c'_η such that the following holds for any $p \geq 3$. Consider any $\theta^*, \sigma, \Sigma \in \mathcal{U}[\eta]$, $\delta \in (0, 1)$, $(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3) > 0$ satisfying $c[\mathbf{a}_2 \|\theta^*\|_0 + \log(\frac{4}{\delta})] \leq m$, and S*

satisfying $\mathbf{S}[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$. Taking

$$\underline{c}_* \geq \sqrt{2}\mathbf{a}_1 + 11\eta^2(\mathbf{a}_3 \vee 1) \sqrt{\log\left(\frac{4e}{\delta}\right)}. \quad (27)$$

we have $\mathbb{P}_{\theta^*, \sigma, \Sigma}^{(0)}[\phi^{(th)}[S, \underline{c}_*] = 1] \leq \delta$ if $\|\theta^*\|_0 \leq k_0$. Besides, $\mathbb{P}_{\theta^*, \sigma, \Sigma}^{(0)}[\phi^{(th)}[S, \underline{c}_*] = 1] \geq 1 - \delta$ if $\|\theta^*\|_0 > k_0$ and

$$d_2^2(\theta^*; \mathbb{B}_0[k_0]) \geq c_\eta \left[1 + \mathbf{a}_3^2 \|\theta^*\|_0 \frac{\log(p)}{m}\right] \left[\mathbf{a}_3^2 \sqrt{\log(4/\delta)} + \underline{c}_*^2\right] \sigma^2 \frac{[\|\theta^*\|_0 - k_0] \log(p)}{m}.$$

If S satisfies $(\mathbf{S}[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3])$, then the test $\phi^{(th)}[S; c]$ with a suitable tuning parameter c has a controlled type I error probability. Besides, its square separation distance over $\mathbb{B}_0[k_0 + \Delta]$ is (up to constants depending on \mathbf{a}_1 and \mathbf{a}_3) of the order of $\Delta \log(p)/n$.

In view of this general result, it suffices to build an estimator \widehat{S} of the support based on $(Y^{(1)}, \mathbf{X}^{(1)})$ that satisfies $(\mathbf{S}[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3])$ for small $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ to get the squared separation distance $\Delta \log(p)/n$.

Unfortunately, the support of the Lasso estimator is only proved to satisfy the first part of property $(\mathbf{S}[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3])$. Its number of false positives is at most of the order of $\|\theta^*\|_0$, see [50]. It turns out that the second part of the property has only recently been proved to be achieved by non-convex penalized estimators, see [24] such as MCP estimator [49].

As in the previous section, we consider the column normalized design $\mathbf{T}^{(1)}$. Given $\beta \in \mathbb{R}^p$ and two tuning parameters $b > 0, \lambda > 0$, the MCP criterion is defined by

$$L(\theta) := \|Y^{(1)} - \mathbf{T}^{(1)}\theta\|_2^2 + \sum_{i=1}^p \rho(|\theta_i|; \lambda); \quad \rho(t; \lambda) = \lambda \int_0^t (1 - x/(\kappa\lambda))_+ dx, \quad (28)$$

where $x_+ = \max(x, 0)$. Local minimizers of the MCP criterion can be efficiently computed using the PLUS Algorithm from [49] or by approximate regularization path by [48]. It turns out that non-convex penalized estimators suffer from less bias than Lasso estimators.

Consider the square-root Lasso estimator (8) $\widehat{\theta}_{SL}$ with $\delta = 1/p$ and the plug-in variance estimator $\widehat{\sigma}_{SL} = \|Y - \mathbf{X}\widehat{\theta}_{SL}\|_2/\sqrt{m}$. Define the tuning parameters

$$\widehat{\lambda}_{MCP} = \underline{c}_\eta^{(MCP)} \widehat{\sigma}_{SL} \sqrt{\log(p)}; \quad \kappa = \underline{c}_\eta'^{(MCP)}$$

for some constants $\underline{c}_\eta^{(MCP)}$ and $\underline{c}_\eta'^{(MCP)}$ whose range of possible values follows from [49] and [24]. The following proposition is a consequence of Corollary 1 in [24] together with Theorem 6 in [49].

Proposition 8. *There exist constants $\underline{c}_\eta^{(MCP)}, c, c_\eta^{(1)} - c_\eta^{(4)}$ such that the following holds for any θ^* with $c_\eta^{(1)} \|\theta^*\|_0 \leq m/\log(p)$. With probability higher than $1 - cp^{-1}$, the support \widehat{S}_{MCP} of any stationary point of the criterion (28) satisfies $\mathbf{S}[c_\eta^{(2)}, c_\eta^{(3)}, c_\eta^{(4)}]$.*

A similar result holds if we use the non-convex SCAD penalty instead of MCP from [24].

Now, we can plug the support estimator \widehat{S}_{MCP} into the test $\phi^{(th)}$ with a suitable constant \underline{c}_* . The following result is a straightforward consequence of Theorems 3 and Proposition 8 and its proof is therefore omitted.

Corollary 1. *There exist constants $c, c_{\eta, \delta}$, and $c'_{\eta, \delta}$ such that the following holds under Condition $(\mathbf{B}(\delta))$. The test $\phi^{(th)}[\widehat{S}_{MCP}; \underline{c}_\eta^{(MCP),*}]$ satisfies $(\mathbf{gP1}[\frac{c}{p^2} + \delta])$ and $(\mathbf{gP2}[\frac{c}{p^2} + \delta])$ over the collections*

$$\mathbb{B}_0[k_0 + \Delta] \cap \left\{ \theta^*, d_2^2[\theta^*; \mathbb{B}_0[k_0]] \geq c_{\eta, \delta} \sigma^2 \frac{\Delta \log(p)}{n} \right\}, \quad (29)$$

for all $1 \leq \Delta \leq c'_{\eta, \delta} n / \log(p)$.

3.2.3 Aggregated tests and summary

Consider some $\delta > 0$. Since the performances of the test $\phi^{(u)}$ are only assessed in the regime $p \leq c_\eta n^2 \log^{-1}(\frac{2}{\delta})$ (c_η is introduced in Proposition 7), we combine the tests $\phi^{(u)}$ and $\phi^{(th)}[\widehat{S}_{MCP}; \underline{\mathcal{L}}_\eta^{(MCP),*}]$ only in that regime. For larger p , we solely use $\phi^{(th)}[\widehat{S}_{MCP}; \underline{\mathcal{L}}_\eta^{(MCP),*}]$. Combining Proposition 7 and Corollary 1 to evaluate the separation distance of the aggregated test and comparing them with the minimax lower bounds of Proposition 6, we obtain the following characterization - note that we assume here that we are in the high dimensional regime, i.e. $p \geq n^{1+\zeta}$ where $\zeta > 0$ is an arbitrarily small absolute constant.

Case 1: $p \leq n^{2-\kappa}$ with an arbitrary but fixed $\kappa \in (0, 1/2)$ and $k_0 \leq \sqrt{p}p^{-\zeta}$.

$$\rho_{g,\gamma}^{*2}[k_0, \Delta] \asymp_{\gamma,\eta,\kappa,\zeta,\varsigma} \begin{cases} \frac{\Delta \log(p)}{n} & \text{if } \Delta \leq \min(\sqrt{p}, k_0)p^{-\zeta} ; \\ \frac{\sqrt{p}}{n} & \text{if } \Delta \geq \sqrt{p} , \end{cases}$$

for $1 \leq \Delta \leq p - k_0$.

Case 2: $p \leq n^{2-\kappa}$ with an arbitrary but fixed $\kappa \in (0, 1/2)$ and $k_0 \geq \sqrt{p}$.

$$\begin{aligned} \rho_{g,\gamma}^{*2}[k_0, \Delta] &\asymp_{\gamma,\eta,\kappa,\zeta} \frac{\Delta \log(p)}{n} && \text{if } \Delta \leq k_0 p^{-\zeta} ; \\ \frac{k_0}{n \log(p)} \lesssim_{\gamma,\eta,\kappa} \rho_{g,\gamma}^{*2}[k_0, \Delta] &\lesssim_{\gamma,\eta,\kappa} \frac{k_0 \log(p)}{n} && \text{if } \Delta \geq k_0 , \end{aligned}$$

for $1 \leq \Delta \leq p - k_0$.

Case 3: $p \geq n^2$. For any k_0 and Δ smaller than $c_\eta n / \log(p)$, we have

$$\rho_{g,\gamma}^{*2}[k_0, \Delta] \asymp_{\gamma,\eta} \frac{\Delta \log(p)}{n} ,$$

whereas the problem become much more difficult for larger Δ or k_0 .

In conclusion, the aggregated test achieves the minimax separation distance except in the regime where $\sqrt{p} \leq k_0 \leq \Delta \leq n$ where there is $\log^2(p)$ gap between the two squared rates.

Summing up our findings, we observe that

- For sparse alternatives (small Δ) - first result in Cases 1 and 2 and result in Case 3 - then the minimax separation distance is analogous to that of signal detection ($k_0 = 0$), i.e. of order $\frac{\Delta \log(p)}{n}$. It would be straightforward to achieve this distance if we had at our disposal a $\sqrt{\log(p)/n}$ l_∞ -consistent estimator of θ^* . However, this is not possible over the class of $\Sigma \in \mathcal{U}(\eta)$ (Σ unknown in this class) [9, 32]. This is why we use a slightly different approach that focuses on selecting most of the relevant features (as in Property $\mathbf{S}[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$ (25)).
- If Δ is large and k_0 is small - second result in Case 1 - then the squared minimax separation distance is of the order of \sqrt{p}/n and is the same as for signal detection ($k_0 = 0$). It is achieved by a U -statistic originally introduced for estimating $\|\theta^*\|_2^2$ when $\Sigma = \mathbf{I}_p$ [22, 47].
- If Δ is large and k_0 is large - second result in Case 2 - then the lower bound on the minimax separation distance reflects the complexity of the null hypothesis. The lower bound is the same as in the independent setting, described in the previous section. The upper bound is based on the same U -statistic as in the previous case. Unfortunately the upper and lower bounds only match up to $\log^2(p)$ factor. In this general setting, we doubt that adapting the Fourier statistic of the previous section is possible and we conjecture that the squared separation distance is actually of the order $k_0 \log(p)/n$.

- Finally, we emphasize that, for Δ large compared to $n/\log(p)$ and $p \geq n^2$, the optimal separation distance is huge (Proposition 6). Without further assumptions, it is therefore almost impossible to test whether θ^* is k_0 -sparse or if θ^* is a dense vector when $p \geq n^2$. This result is in sharp contrast with the independent setting.

3.2.4 An alternative variable selection procedure

In the previous section, we established that the test $\phi^{(th)}$ applied to the support \widehat{S}_{MCP} estimated by the MCP estimator achieves the square separation rate $\Delta \log(p)/n$. Here, we introduce an alternative to the concave penalized estimator MCP based on simple iterations of the thresholded square-root Lasso.

Starting from $\widehat{S}_0 = \emptyset$, the algorithm builds a subset \widehat{S}_t of variables iteratively from a subset \widehat{S}_{t-1} of variables. It is done by applying a thresholded square-root Lasso to the data projected on the orthogonal of the variables in \widehat{S}_{t-1} . Then, \widehat{S}_t is the concatenation of \widehat{S}_{t-1} and the variables selected by the thresholded square-root Lasso. The procedure stops after approximately $\log(n)$ iterations, and returns the current subset. The general idea is to iteratively remove non-zero coordinates of θ^* so that the projected square-root Lasso estimator is less perturbed by large coordinates of θ^* .

We need to introduce some notation. Define $T = \lfloor \log_2(n) \rfloor + 1$. Assume without loss of generality that $m/T = n/(2T)$ is an integer. We divide the sample $(Y^{(1)}, \mathbf{X}^{(1)})$ into T subsamples $\{(\underline{Y}^{(t)}, \underline{\mathbf{X}}^{(t)})\}$ of size m/T . Given a $r \times d$ matrix \mathbf{M} and some subset $S \subset \{1, \dots, p\}$, we write \mathbf{M}_S for the $r \times d$ matrix defined by $(\mathbf{M}_S)_{i,j} = \mathbf{M}_{i,j} \mathbf{1}_{\{j \in S\}}$. Given S and any $1 \leq t \leq T$, define the subspace $V[S, \underline{\mathbf{X}}^{(t)}] = \text{vect}(\underline{\mathbf{X}}_S^{(t)})$ of $\mathbb{R}^{m/T}$ and an $(m/T - \dim(V[S, \underline{\mathbf{X}}^{(t)}])) \times m/T$ matrix $\underline{\mathbf{P}}_{t,S}^\perp$ (measurable with respect to $\underline{\mathbf{X}}_S^{(t)}$) whose corresponding linear application is null on $V[S, \underline{\mathbf{X}}^{(t)}]$ and maps isometrically the orthogonal of $V[S, \underline{\mathbf{X}}^{(t)}]$ to $\mathbb{R}^{m/T - \dim(V[S, \underline{\mathbf{X}}^{(t)}])}$.

Next, we define the **Thresholded square-root Lasso estimator**. Let $\underline{m} > 0$ and let $\delta > 0$. Given a $\underline{m} \times p$ matrix $\underline{\mathbf{X}}$ and a size \underline{m} vector \underline{Y} , we write $\underline{\mathbf{X}}_c$ for the subdesign matrix of $\underline{\mathbf{X}}$ where its null rows have been removed. Then, $\widehat{\theta}_{SL}(\underline{\mathbf{X}}, \underline{Y})$ stands for the square-root Lasso estimator (see Equation (8)) of $(\underline{\mathbf{X}}_c, \underline{Y})$ with parameter $\lambda = 2\sqrt{\widehat{\Phi}^{-1}(\delta/(2p))}$. For the purpose of notation, we consider that $\widehat{\theta}_{SL}(\underline{\mathbf{X}}, \underline{Y}) \in \mathbb{R}^p$ and that its entries $\widehat{\theta}_{SL}(\underline{\mathbf{X}}, \underline{Y})$ corresponding to null rows of $\underline{\mathbf{X}}$ are equal to zero. Using the plug-in variance estimator $\widehat{\sigma}^2 = \|\underline{Y} - \underline{\mathbf{X}}[\widehat{\theta}_{SL}(\underline{\mathbf{X}}, \underline{Y})]\|_2^2/\underline{m}$, we define the thresholding modification $\widehat{\theta}_{SL,t}(\underline{\mathbf{X}}, \underline{Y})$ of $\widehat{\theta}_{SL}(\underline{\mathbf{X}}, \underline{Y})$ such that

$$[\widehat{\theta}_{SL,t}(\underline{\mathbf{X}}, \underline{Y})]_i = [\widehat{\theta}_{SL}(\underline{\mathbf{X}}, \underline{Y})]_i \mathbf{1}\left\{ |[\widehat{\theta}_{SL}(\underline{\mathbf{X}}, \underline{Y})]_i| \geq \underline{c}_\eta^{(SL)} \widehat{\sigma} \frac{8}{3} \sqrt{\frac{\log(p/\delta)}{\underline{m}}} \right\}, \quad i = 1, \dots, p,$$

where the constant $\underline{c}_\eta^{(SL)}$ is introduced in Lemma 1.

The set $\widehat{S}^{(ith)}$ is constructed as follows. We start with the empty support $\widehat{S}_0 = \emptyset$. At each step $t = 1, \dots, T$, we project both $\underline{\mathbf{X}}^{(t)}$ and $\underline{Y}^{(t)}$ along the space $V[\widehat{S}_{t-1}, \underline{\mathbf{X}}^{(t)}]$ spanned by the variables in \widehat{S}_{t-1} . Then, we apply thresholded square-root Lasso to these projected data to select new variables. Finally, $\widehat{S}^{(ith)}$ is the last set \widehat{S}_T .

Algorithm 1 Iterative construction of a set of relevant coordinates

Input: $\eta, \mathbf{X}, \mathbf{Y}$ $\{\eta$ is required to compute $\underline{\mathcal{L}}_\eta^{(SL)}$ in the definition of $\widehat{\theta}_{(SL,t)}\}$

1: $\widehat{S}_0 = \emptyset$ {Initialization of the set of relevant coordinates}

2: **for** $t = 1, \dots, T$, **do**

3: $\widehat{S}_t = \widehat{S}_{t-1} \cup \mathcal{S}(\widehat{\theta}_{(SL,t)}[\underline{\mathbf{\Pi}}_{t,\widehat{S}_{t-1}}^\perp \mathbf{X}^{(t)}, \underline{\mathbf{\Pi}}_{t,\widehat{S}_{t-1}}^\perp \mathbf{Y}^{(t)}])$ {We re-compute at each step the thresholded square-root Lasso } on data that have been projected on the orthogonal of previous supports}

4: **end for**

Output: $\widehat{S}^{(ith)} = \widehat{S}_T$

Theorem 4. *There exist constants $\underline{\mathcal{L}}_\eta^{(ith)}$ and c'_η such that the following holds for any $\sigma > 0$, $\Sigma \in \mathcal{U}(\eta)$ and any θ^* satisfying*

$$c'_\eta \|\theta^*\|_0 \leq \frac{n}{\log^2(n) \log(p)}. \quad (30)$$

With probability higher than $1 - Tp^{-2}$, the estimator $\widehat{S}^{(ith)}$ satisfies $\mathbf{S}[\underline{\mathcal{L}}_\eta^{(ith)} \sqrt{T}, 2T, \underline{\mathcal{L}}_\eta^{(ith)} \sqrt{T/2}]$.

It turns out that $\widehat{S}^{(ith)}$ satisfies the desired property $\mathcal{S}[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$ but with \mathbf{a}_1 and \mathbf{a}_3 that are logarithmically large. As a consequence, the squared separation distance of the corresponding test $\phi^{(th)}$ with $\widehat{S}^{(ith)}$ is of order $\Delta \log(p) \log(n)/n$, which is optimal up to an additional $\log(n)$ term in the regime $\Delta \leq k_0 \wedge \sqrt{p}$.

4 Discussion

In this section, we briefly discuss several related problems.

4.1 Low-dimensional problems

Although some of our results are valid in a low-dimensional setting, we focused our attention on pinpointing the minimax separation distance in a high-dimensional regime $p \geq n^{1+\zeta}$ which is arguably the most interesting one. Let us briefly discuss the low dimensional regime $p \leq n/2$. In the independent setting, the main difference is that the $n^{-1/2}$ rate can be improved to $\sqrt{p}/n + k_0 \log(p)/n$ by considering the ordinary least-square estimator and computing its l_2 -norm when its k_0 largest entries are removed. In the general setting, we can recover similar upper bounds as in Section 3, but with much simpler procedures based on the ordinary least-squares estimator.

Between these two regimes, the medium-dimensional case where p and n are of the same order is technically challenging. Our upper bounds and lower bounds only match up to polylogarithmic factor. Deriving the sharp minimax separation distance requires further work.

4.2 Sparse inverse covariance matrices Σ^{-1} and debiased Lasso

Consider an intermediary setting where both σ and Σ are unknown but Σ^{-1} is also restricted to have less than $\sqrt{n}/\log(p)$ non-zero entries on each rows. In this setting, the minimax lower bounds of Proposition 6 in the general setting turn out to be still valid. Indeed, the proof of Proposition 6 holds in the simpler setting where $\Sigma = \mathbf{I}_p$ and σ is unknown. As in the general setting, the upper bound $k_0 \log(p)/n + \sqrt{p}/n$ is achieved by the polynomial time U -statistic of Section 3.2.1. In contrast, achieving the $\sqrt{\Delta \log(p)/n}$ separation distance in the small Δ regime is now much easier than in the general setting. Whereas we introduced a refitted least-square estimator combined with the non-convex MCP regularized estimator, one can now alternatively rely on the debiased Lasso

method [31, 32, 44, 51] to obtain a $\sqrt{\log(p)/n}$ l_∞ -consistent estimator of θ^* and then simply count the number of its large entries. This was already done in [30] as discussed previously.

4.3 Know Σ and unknown σ^2 .

Consider the intermediate scenario where $\Sigma = \mathbf{I}_p$ is the identity matrix, but σ^2 is unknown. As explained in the previous subsection, the minimax lower bounds of Proposition 6 stated for the general setting are still valid in this intermediate scenario. Obviously, we can also apply the testing procedures of Section 3.

However, in the general scenario, our lower and upper bounds are only matching up to a $\log^2(p)$ factor in the large k_0 , large Δ setting. More specifically, when $k_0 \geq p^{1/2+\zeta}$ (for some $\zeta > 0$) and $\Delta \geq k_0$, the lower bound of Proposition 6 is of order $k_0/[n \log(p)]$ whereas Proposition 7 provides an upper bound of the order of $k_0 \log(p)$.

It turns out that, in this intermediate scenario, the gap is easily closed by adapting the Fourier-based test $\phi^{(f)}$ and $\phi^{(i)}$ introduced in Section 2. Indeed, the only place where the knowledge of σ is necessary in these two tests is in the definition of the pre-estimator $\bar{\theta}_{\mathbf{I}}$ which is a thresholded version of $\tilde{\theta}_{\mathbf{I}}$ (9). If we replace σ in this threshold by the plug-in estimator of the variance based on the square-root Lasso and if we increase some constants, this modification of the tests $\phi^{(f)}$ and $\phi^{(i)}$ does not depend anymore on σ . Besides, one can easily check that (up to some changes in the numerical constants) Propositions 4 and 5 are still valid for these tests.

4.4 Unknown Σ and known σ^2 .

In this case, we can improve the upper bounds of the general case by adapting the test $\phi^{(\chi)}$ from Section 2. Indeed, the statistic $Z_{\chi}(\hat{R}_{k_0})$ is now centered on $\|\Sigma^{1/2}(\theta^* - \tilde{\theta}_{SL, k_0})\|_2^2 \geq \eta^{-1} \|\theta^* - \tilde{\theta}_{SL, k_0}\|_2^2$ on the class $\mathcal{U}(\eta)$ of Σ . Hence, the corresponding test achieves a squared separation distance of the order of $n^{-1/2} + k_0 \log(p)/n$. The main difference with the independent case is that we are not able to adapt the Fourier-based test $\phi^{(f)}$ and $\phi^{(i)}$ to unknown Σ . In regimes where $p^{1/2+\kappa} \leq k_0 \leq c_\gamma \frac{n}{\log(p)}$ and $\Delta \geq k_0$, there is therefore a $\log^2(p)$ gap between our upper and lower bounds.

5 Proofs of the minimax upper bounds

5.1 Some results on the square-root Lasso and a simple debiased Lasso

We start with a few probability bounds for the square-root Lasso $\hat{\theta}_{SL}$ and its thresholded modification $\tilde{\theta}_{SL, k_0}$ where only the k_0 largest values of $\hat{\theta}_{SL}$ are not set to 0. They almost follow straightforwardly from earlier results [4, 36, 40, 42]. As we shall apply this lemma in different contexts, we reintroduce the setting here. We consider a $m \times q$ linear regression model $Y = \mathbf{X}\theta^* + \sigma\epsilon$ with $\epsilon \sim \mathcal{N}(0, \mathbf{I}_m)$ and where the rows of \mathbf{X} are independent and follow a centered normal distribution with common covariance matrix Σ . The $m \times q$ matrix \mathbf{T} is the column normalized version of \mathbf{X} i.e.

$$\mathbf{T} = \left(\mathbf{X}_{\cdot,1}/\|\mathbf{X}_{\cdot,1}\|_2, \dots, \mathbf{X}_{\cdot,q}/\|\mathbf{X}_{\cdot,q}\|_2 \right).$$

We take

$$\lambda_{\delta, m, q} = 2\sqrt{\Phi^{-1}(\delta/(4q))}, \quad (31)$$

and consider the square-root Lasso estimator [4, 42],

$$\hat{\theta}_{SL, N} \in \arg \min \|Y - \mathbf{T}\theta\|_2 + \lambda_{\delta, m, q} \|\theta\|_1, \quad (32)$$

Then, define $(\widehat{\theta}_{SL})$ as $(\widehat{\theta}_{SL})_i = (\widehat{\theta}_{SL,N})_i / \|\mathbf{X}_{\cdot,i}\|_2$ for any $i = 1, \dots, q$ and $\widetilde{\theta}_{SL,k_0} = \arg \min_{\theta \in \mathbb{B}_0[k_0]} \|\theta - \widehat{\theta}_{SL}\|_2^2$. The plug-in variance estimator is $\widehat{\sigma}_{SL} = \|Y - \mathbf{X}\widehat{\theta}_{SL}\|_2 / \sqrt{m}$.

Lemma 1. *Fix any $0 < \delta \leq 1/2$ and any $\eta \geq 1$. There exist constants $\underline{c}_\eta^{(SL)}$ and $\underline{c}_\eta^{(SL),2}$ such that the following holds. Let k_{\max} be the largest integer such that*

$$\underline{c}_\eta^{(SL),2} [k_{\max} \log(q/\delta) + \log(1/\delta) \log(q)] \leq m ,$$

For any $\sigma > 0$, $\Sigma \in \mathcal{U}[\eta]$ and θ^ with $\|\theta^*\|_0 \leq k_{\max}$, there exists an event \mathcal{E} of probability higher than $1 - \delta$, such that*

$$\begin{aligned} \|\widehat{\theta}_{SL} - \theta^*\|_2^2 &\leq \underline{c}_\eta^{(SL)} \sigma^2 \frac{\|\theta^*\|_0 \vee 1}{m} \log\left(\frac{q}{\delta}\right) ; \\ \|\widetilde{\theta}_{SL,k_0} - \theta^*\|_2^2 &\leq \underline{c}_\eta^{(SL)} \left[\sigma^2 \frac{\|\theta^*\|_0 \vee 1}{m} \log\left(\frac{q}{\delta}\right) + d_2^2[\theta^*; \mathbb{B}_0[k_0]] \right] ; \\ \frac{3}{4} &\leq \frac{\widehat{\sigma}_{SL}}{\sigma} \leq \frac{5}{4} . \end{aligned}$$

Proof of Lemma 1. We first argue that the design matrix \mathbf{T} satisfies the compatibility property (see [36, 42]) with any set of size less than k_{\max} and constant depending on η . Indeed, Corollary 1 in [40] enforces that this property holds with probability higher than $1 - qe^{-cm} \geq 1 - \delta/2$. Then, we are in position to apply Theorem 1 in [42], which implies that $\widehat{\sigma}_{SL}/\sigma$ belongs to $[3/4, 5/4]$ and that

$$\|\widehat{\theta}_{SL} - \theta^*\|_2^2 \leq \underline{c}_\eta^{(SL)} \sigma^2 \frac{\|\theta^*\|_0 \vee 1}{m} \log\left(\frac{q}{\delta}\right) .$$

The second result of the lemma is a consequence of the first result. Denote S_1 (resp. S_2) the subset of the k_0 largest entries of $\widehat{\theta}_{SL}$ (resp. θ^*). From the definition of $\widetilde{\theta}_{SL,k_0}$, we deduce that

$$\begin{aligned} \|\widetilde{\theta}_{SL,k_0} - \theta^*\|_2^2 &\leq 2\|\widehat{\theta}_{SL} - \theta^*\|_2^2 + 2 \sum_{i \in \overline{S_1}} (\widehat{\theta}_{SL})_i^2 \\ &\leq 2\|\widehat{\theta}_{SL} - \theta^*\|_2^2 + 2 \sum_{i \in \overline{S_2}} (\widehat{\theta}_{SL})_i^2 \\ &\leq 6\|\widehat{\theta}_{SL} - \theta^*\|_2^2 + 4 \sum_{i \in \overline{S_2}} \theta_i^{*2} = 6\|\widehat{\theta}_{SL} - \theta^*\|_2^2 + 4d_2^2[\theta^*; \mathbb{B}_0[k_0]] . \end{aligned}$$

The result follows. \square

5.2 Analysis of the tests $\phi^{(t)}$, $\phi^{(\chi)}$, and $\phi^{(u)}$

5.2.1 Proof of Proposition 2 (Test $\phi^{(t)}$)

We start with a l_∞ error bound on $\widetilde{\theta}_{\mathbf{I}}$.

Lemma 2. *There exists a constant c such that the following holds under $(\mathbf{A}[\alpha \wedge \delta])$. For any $\theta^* \in \mathbb{R}^p$ with $\|\theta^*\|_0 \leq k_{\max}$ (with k_{\max} as in Lemma 1), we have*

$$\|\theta^* - \widetilde{\theta}_{\mathbf{I}}\|_\infty \leq c\sigma \sqrt{\frac{\log(p/\alpha)}{n}} ,$$

with probability higher than $1 - \delta - \alpha$. Besides, for any $\theta^ \in \mathbb{R}^p$, we have*

$$\|\theta^* - \widetilde{\theta}_{\mathbf{I}}\|_\infty \mathbf{1}_{\|\widehat{\theta}_{SL} - \theta^*\|_2^2 \leq 2\sigma^2} \leq c\sigma \sqrt{\frac{\log(p/\alpha)}{n}} ,$$

with probability higher than $1 - \alpha$.

From Lemma 2, we derive that with probability higher than $1 - \delta - \alpha$, we have $\|\theta^* - \tilde{\theta}_{\mathbf{I}}\|_{\infty} \leq c\sigma\sqrt{\frac{\log(p/\alpha)}{n}}$. Setting $\underline{c}^{(t)}$ as c in Lemma 2, we derive that the test $\phi^{(t)}$ has a type I error probability less or equal to $\alpha + \delta$. Now consider a vector $\theta^* \in \mathbb{B}_0[k_0 + \Delta]$ such that $|\theta_{(k_0+1)}^*| \geq 2.1\underline{c}^{(t)}\sigma\sqrt{\log[p/(\alpha \wedge \beta)]/n}$. From Lemma 2, we deduce that, with probability higher than $1 - \alpha - \beta$,

$$|(\tilde{\theta}_{\mathbf{I}})_{(k_0+1)}| \geq |\theta_{(k_0+1)}^*| - \|(\tilde{\theta}_{\mathbf{I}}) - \theta^*\|_{\infty} \geq \underline{c}^{(t)}\sigma(2.1\sqrt{\log(p/(\alpha \wedge \beta))} - \sqrt{\log(p/\beta)})\frac{\sigma}{\sqrt{n}} > \underline{c}^{(t)}\sigma\sqrt{\frac{\log(p/\alpha)}{n}},$$

and the test $\phi^{(t)}$ therefore rejects the null hypothesis, which concludes the proof.

Proof of Lemma 2. Set $\hat{\gamma} = \theta^* - \hat{\theta}_{SL}$. If $\|\theta^*\|_0 \leq k_{\max}$, the conditions of Lemma 1 are satisfied and it follows from this lemma that

$$\|\hat{\gamma}\|_2^2 \leq c\sigma^2(\|\theta^*\|_0 \vee 1)\frac{\log(p/\delta)}{m} \leq c'\sigma^2, \quad (33)$$

with $\mathbb{P}_{\theta^*, \sigma}^{(1)}$ probability higher than $1 - \delta$. In the second result of Lemma 2, we restrict ourselves to the case $\|\hat{\gamma}\|_2^2 \leq 2\sigma^2$. Hence, it suffices to prove that, conditionally to $\hat{\gamma}$ satisfying $\|\hat{\gamma}\|_2^2 \leq (2 \vee c')\sigma^2$, we have $\|\theta^* - \tilde{\theta}_{\mathbf{I}}\|_{\infty} \leq c_1\sigma\sqrt{\frac{\log(p/\alpha)}{n}}$ with probability higher than $1 - \alpha$.

Write Z the statistic defined by $Z = \theta^* - \hat{\theta}_{SL} - \frac{1}{m}\mathbf{X}^{(2)T}(Y^{(2)} - \mathbf{X}^{(2)}\hat{\theta}_{SL})$. Also, define $\hat{\Sigma} = \frac{1}{m}\mathbf{X}^{(2)T}\mathbf{X}^{(2)}$ and $\hat{\Gamma}$ its diagonal part. We have

$$\begin{aligned} \|Z\|_{\infty} &= \left\| \hat{\gamma} - \frac{1}{m}\mathbf{X}^{(2)T}\mathbf{X}^{(2)}\hat{\gamma} - \frac{\sigma}{m}\mathbf{X}^{(2)T}\epsilon^{(2)} \right\|_{\infty} \\ &\leq \left\| \hat{\gamma} - \frac{1}{m}\mathbf{X}^{(2)T}\mathbf{X}^{(2)}\hat{\gamma} \right\|_{\infty} + \left\| \frac{\sigma}{m}\mathbf{X}^{(2)T}\epsilon^{(2)} \right\|_{\infty} \\ &\leq \|(\mathbf{I}_m - \hat{\Gamma})\hat{\gamma}\|_{\infty} + \|(\hat{\Sigma} - \hat{\Gamma})\hat{\gamma}\|_{\infty} + \left\| \frac{\sigma}{m}\mathbf{X}^{(2)T}\epsilon^{(2)} \right\|_{\infty} = A_1 + A_2 + A_3. \end{aligned}$$

We control each of these three quantities independently.

Lemma 3. Let \mathbf{Q} be a $d \times d$ symmetric matrix and let $G \sim \mathcal{N}(0, \mathbf{I}_d)$. Define $S = G^T\mathbf{Q}G$. For any $t > 0$, one has

$$S \leq \text{Tr}(\mathbf{Q}) + 2\|\mathbf{Q}\|_F t + 2\|\mathbf{Q}\|_{op} t,$$

with probability higher than $1 - e^{-t}$. Here, $\|\mathbf{Q}\|_F$ and $\|\mathbf{Q}\|_{op}$ respectively correspond to the Frobenius and operator norm of \mathbf{Q} .

This result is a slight extension of Lemma 1 in [37] (that requires \mathbf{Q} to be positive). The extension to general symmetric matrices proceeds from the same arguments and we omit the proof.

Let us first control A_3 . Each of the p entries of $\sigma m^{-1}\mathbf{X}^{(2)T}\epsilon^{(2)}$ is distributed as a quadratic form of $2m$ standard normal random variables. The corresponding matrix \mathbf{Q} satisfies $\text{tr}(\mathbf{Q}) = 0$ and $\|\mathbf{Q}\|_F^2 = \sigma^2/(2m)$. Since $\|\mathbf{Q}\|_{op} \leq \|\mathbf{Q}\|_F$, it follows from the above lemma together with an union bound that

$$A_3 \leq 4\sigma\sqrt{\frac{\log(6p/\alpha)}{2m}}. \quad (34)$$

with $\mathbb{P}_{\theta^*, \sigma}^{(2)}$ probability higher than $1 - \alpha/3$. As for A_1 and A_2 , we first work conditionally to $\hat{\gamma}$. Fix $i \in \{1, \dots, p\}$, $\hat{\Gamma}_{ii}$ is distributed as quadratic form of m standard normal variable and the

corresponding matrix \mathbf{Q} satisfies $\text{tr}(\mathbf{Q}) = 1$, $\|\mathbf{Q}\|_F = m^{-1/2}$ and $\|\mathbf{Q}\|_{op} \leq 1/m$. It then follows from Lemma 3, that conditionally to $\hat{\gamma}$,

$$A_1 \leq 2(1 + m^{-1/2})\sqrt{\frac{\log(6p/\alpha)}{m}}\|\hat{\gamma}\|_\infty, \quad (35)$$

with $\mathbb{P}_{\theta^*, \sigma}^{(2)}$ probability higher than $1 - \alpha/3$. As for A_2 , observe that, conditionally to $\hat{\gamma}$, $[(\hat{\Sigma} - \hat{\Gamma})\hat{\gamma}]_j = \frac{1}{m} \sum_i \mathbf{X}_{i,j} \sum_{j' \neq j} \mathbf{X}_{i,j'} \hat{\gamma}_{j'}$ is distributed as $(\sum_{j' \neq j} \hat{\gamma}_{j'}^2)^{1/2}/m \sum_{q=1}^m U_q U'_q$ where the U_q 's and U'_q 's are independent standard normal random variables. Again, we deduce from Lemma 3 that, conditionally to $\hat{\gamma}$,

$$A_2 \leq 4\sqrt{\frac{\log(6p/\alpha)}{2m}} \max_i \left(\sum_{j \neq i} \hat{\gamma}_j^2 \right)^{1/2} \leq 4\|\hat{\gamma}\|_2 \sqrt{\frac{\log(6p/\alpha)}{2m}}, \quad (36)$$

with $\mathbb{P}_{\theta^*, \sigma}^{(2)}$ probability higher than $1 - \alpha/6$. Finally, we gather (33) with (34–36) to conclude that there exists $c > 0$ such that

$$\|Z\|_\infty \leq c\sigma \sqrt{\frac{\log(p/\alpha)}{m}},$$

with probability larger than $1 - \delta - \alpha$. □

5.2.2 Proof of Proposition 3 (Test $\phi^{(x)}$)

We first state the following lemma that characterizes the deviations of $Z_\chi[R]$.

Lemma 4. *For any $t > 0$, any $\theta^* \in \mathbb{R}^p$, any $\sigma > 0$, and any fixed θ , we have for, $\hat{R} = Y^{(2)} - \mathbf{X}^{(2)}\theta$,*

$$\left| Z_\chi(\hat{R}) - \frac{\|\theta^* - \theta\|_2^2}{\sigma^2} \right| \leq 2 \left[1 + \frac{\|\theta^* - \theta\|_2^2}{\sigma^2} \right] \left[\sqrt{\frac{t}{m}} + \frac{t}{m} \right], \quad (37)$$

with $\mathbb{P}_{\theta^*, \sigma}^{(2)}$ probability higher than $1 - e^{-t}$.

Proof of Lemma 4. We have

$$\begin{aligned} \hat{R}_i &= (Y^{(2)} - \mathbf{X}^{(2)}\theta)_i = \epsilon_i^{(2)} + (\mathbf{X}^{(2)}(\theta^* - \theta))_i \\ &= \epsilon_i^{(2)} + \sum_j \mathbf{X}_{i,j}^{(2)}(\theta_j^* - \theta_j). \end{aligned}$$

Hence, $\hat{R}_i \sim \mathcal{N}(0, \sigma^2 + \|\theta^* - \theta\|_2^2)$ and these variables are independent from each other. So the random variable $\|\hat{R}\|_2^2[\sigma^2 + \|\theta^* - \theta\|_2^2]^{-1}$ follows a χ^2 distribution with n degrees of freedom. To prove the result, we only have to apply Lemma 3 with $\mathbf{Q} = \mathbf{I}_p$. □

First assume that θ^* belongs to $\mathbb{B}_0[k_0]$. With $\mathbb{P}_{\theta^*, \sigma}^{(2)}$ probability higher than $1 - \alpha$, we have

$$\begin{aligned} Z_\chi(\hat{R}_{k_0}) &\leq \frac{\|\theta^* - \tilde{\theta}_{SL, k_0}\|_2^2}{\sigma^2} + 2 \left[1 + \frac{\|\theta^* - \tilde{\theta}_{SL, k_0}\|_2^2}{\sigma^2} \right] \left[\sqrt{\frac{\log(1/\alpha)}{m}} + \frac{\log(1/\alpha)}{m} \right] \\ &\leq \frac{\|\theta^* - \tilde{\theta}_{SL, k_0}\|_2^2}{\sigma^2} + c \left[1 + \frac{\|\theta^* - \tilde{\theta}_{SL, k_0}\|_2^2}{\sigma^2} \right] \sqrt{\frac{\log(1/\alpha)}{m}}, \end{aligned}$$

where we used Condition $(\mathbf{A}[\alpha \wedge \delta])$. Then, we apply Lemma 1 to control $\|\theta^* - \tilde{\theta}_{SL, k_0}\|_2^2$ with probability higher than $1 - \delta$. With probability higher than $1 - \alpha - \delta$, we get

$$Z_\chi(\widehat{R}_{k_0}) \leq c \left[(k_0 \vee 1) \frac{\log(p/\delta)}{m} + \sqrt{\frac{\log(1/\alpha)}{m}} \right],$$

so that choosing the constant $\underline{c}^{(\chi)}$ large enough leads to $(\mathbf{P1}[\alpha + \delta])$.

Now assume that $d_2(\theta^*; \mathbb{B}_0[k_0]) > 0$. Since $\tilde{\theta}_{SL, k_0}$ is k_0 -sparse, it follows that $\|\theta^* - \tilde{\theta}_{SL, k_0}\|_2^2 \geq d_2^2[\theta^*; \mathbb{B}_0[k_0]]$. Then, Lemma 4 enforces that, for $\log(1/\beta)$ small enough compared to n (which is ensured by Condition $(\mathbf{A}[\alpha \wedge \beta \wedge \delta])$), one has

$$\begin{aligned} Z_\chi(\widehat{R}_{k_0}) &\geq \frac{d_2^2[\theta^*; \mathbb{B}_0[k_0]]}{\sigma^2} \left[1 - 4\sqrt{\frac{\log(1/\beta)}{m}} \right] - 4\sqrt{\frac{\log(1/\beta)}{m}} \\ &\geq \frac{d_2^2[\theta^*; \mathbb{B}_0[k_0]]}{2\sigma^2} - 4\sqrt{\frac{\log(1/\beta)}{m}}, \end{aligned}$$

with $\mathbb{P}_{\theta^*, \sigma}^{(2)}$ probability larger than $1 - \beta$. As a consequence, under condition (11) with a constant c large enough, the type II error probability is less than β .

5.2.3 Proof of Proposition 7 (test $\phi^{(u)}$)

The following lemma is borrowed from Theorem 2.1 in [47].

Lemma 5. *There exist numerical constants $c > 0$ and $c' > 0$ such that the following holds. Assume that $p \geq m$. Consider any $\theta^* \in \mathbb{R}^p$, any $\sigma > 0$, and any $\Sigma \in \mathcal{U}(\eta)$. Given any estimator $\widehat{\theta}$ based on the subsample $(Y^{(1)}, \mathbf{X}^{(1)})$, define $\widehat{R} = Y^{(0)} - \mathbf{X}^{(0)}\widehat{\theta}$. We have, conditionally on $(\mathbf{X}^{(1)}, Y^{(1)})$,*

$$\mathbb{P}_{\theta^*, \sigma, \Sigma}^{(0)} \left[\left| Z^{(u)} - \frac{(\theta^* - \widehat{\theta})^T \Sigma^2 (\theta^* - \widehat{\theta})}{\sigma^2 + (\theta^* - \widehat{\theta})^T \Sigma (\theta^* - \widehat{\theta})} \right| \leq c\eta \frac{\sqrt{pt}}{m} \right] \geq 1 - c'e^{-t}, \quad (38)$$

for all $t \leq n^{1/3}$.

First, assume that θ^* belongs to $\mathbb{B}_0[k_0]$. Since Condition $(\mathbf{B}[\alpha, \beta])$ is satisfied with a constant $\underline{c}_{\mathbf{B}}^\eta$ large enough, we can apply (38). With probability higher than $1 - \alpha$, one has

$$\begin{aligned} Z^{(u)} &\leq \frac{(\theta^* - \tilde{\theta}_{SL, k_0})^T \Sigma^2 (\theta^* - \tilde{\theta}_{SL, k_0})}{\sigma^2 + (\theta^* - \tilde{\theta}_{SL, k_0})^T \Sigma (\theta^* - \tilde{\theta}_{SL, k_0})} + c\eta \frac{\sqrt{p \log(c'/\alpha)}}{m} \\ &\leq \eta^2 \frac{\|\theta^* - \tilde{\theta}_{SL, k_0}\|_2^2}{\sigma^2} + c\eta \frac{\sqrt{p \log(c'/\alpha)}}{m}. \end{aligned}$$

Then, we use Lemma 1 to conclude that

$$Z^{(u)} \leq c_\eta \left[(k_0 \vee 1) \frac{\log(p/\delta)}{m} + \frac{\sqrt{p \log(c'/\alpha)}}{m} \right],$$

with probability higher than $1 - \alpha - \delta$. Setting the constant $\underline{c}_\eta^{(u)}$ small enough, we conclude that the type I error probability of $\phi^{(u)}$ is less than $\alpha + \delta$.

Now assume that $d_2(\theta^*; \mathbb{B}_0[k_0]) > 0$. Since $\tilde{\theta}_{SL, k_0}$ is k_0 -sparse, it follows that $\|\theta^* - \tilde{\theta}_{SL, k_0}\|_2^2 \geq d_2^2[\theta^*; \mathbb{B}_0[k_0]]$. Then, Lemma 5 enforces that, for $\log(1/\beta)$ small enough compared to n , with probability higher than $1 - \beta$, one has

$$\begin{aligned} Z^{(u)} &\geq \frac{(\theta^* - \tilde{\theta}_{SL, k_0})^T \Sigma^2 (\theta^* - \tilde{\theta}_{SL, k_0})}{\sigma^2 + (\theta^* - \tilde{\theta}_{SL, k_0})^T \Sigma (\theta^* - \tilde{\theta}_{SL, k_0})} - c\eta \frac{\sqrt{p \log(c'/\beta)}}{m} \\ &\geq \eta^{-1} \frac{(\theta^* - \tilde{\theta}_{SL, k_0})^T \Sigma (\theta^* - \tilde{\theta}_{SL, k_0})}{\sigma^2 + (\theta^* - \tilde{\theta}_{SL, k_0})^T \Sigma (\theta^* - \tilde{\theta}_{SL, k_0})} - c\eta \frac{\sqrt{p \log(c'/\beta)}}{m} \\ &\geq \frac{1}{2\eta} \left[\frac{(\theta^* - \tilde{\theta}_{SL, k_0})^T \Sigma (\theta^* - \tilde{\theta}_{SL, k_0})}{\sigma^2} \wedge 1 \right] - c\eta \frac{\sqrt{p \log(c'/\beta)}}{m} \\ &\geq \frac{1}{2\eta^2} \left[\frac{d_2^2[\theta^*; \mathbb{B}_0[k_0]]}{\sigma^2} \wedge 1 \right] - c\eta \frac{\sqrt{p \log(c'/\beta)}}{m}, \end{aligned}$$

where we used in the second and fourth line that Σ belongs $\mathcal{U}(\eta)$. Now assume that $d_2(\theta^*; \mathbb{B}_0[k_0])$ is large enough so that Condition (24) is satisfied. Choosing the constant c'_η in (24) large enough and the constant c_η small enough, it then follows that the type II error probability is smaller than β .

5.3 Analysis of $\phi^{(f)}$ and $\phi^{(i)}$ (Propositions 4 and 5)

In the proofs of this subsection, we set

$$\theta = \theta^* - \bar{\theta}_{\mathbf{I}}.$$

To alleviate the notation, and since $\bar{\theta}_{\mathbf{I}}$ only depends on the first two subsamples, θ can be considered as fixed when we condition to these subsamples. To simplify the notation, we respectively write henceforth Y and \mathbf{X} for $\bar{Y}^{(3)}$ and $\mathbf{X}^{(3)}$ and work conditionally to θ . For any $1 \leq i \leq m$ and $1 \leq j \leq p$, we have $\text{Var}(Y_i) = \sigma^2 + \|\theta\|_2^2$ and $\text{Cov}(Y_i, \mathbf{X}_{i,j}) = \theta_j$. Hence, we have

$$\mathbf{X}_{i,\cdot} | (Y, \theta) \sim \mathcal{N} \left(\frac{Y_i \theta}{\sigma^2 + \|\theta\|_2^2}, \left(\mathbf{I}_p - \frac{\theta \theta^T}{\sigma^2 + \|\theta\|_2^2} \right) \right),$$

and since the $\mathbf{X}_{i,\cdot} | (Y, \theta)$ are independent, we have

$$\mathbf{X}^T Y | (Y, \theta) \sim \mathcal{N} \left(\frac{\|Y\|_2^2 \theta}{\sigma^2 + \|\theta\|_2^2}, \|Y\|_2 \left(\mathbf{I}_p - \frac{\theta \theta^T}{\sigma^2 + \|\theta\|_2^2} \right) \right).$$

For $W = \mathbf{X}^T Y$, it holds that

$$\frac{W}{\|Y\|_2} \Big| (\|Y\|_2, \theta) \sim \mathcal{N} \left(v, \mathbf{I}_p - \frac{\theta \theta^T}{\sigma^2 + \|\theta\|_2^2} \right); \quad \text{with } v = \frac{\theta \|Y\|_2}{\sigma^2 + \|\theta\|_2^2}. \quad (39)$$

As a consequence, given $\|Y\|_2$ and θ , $\frac{W}{\|Y\|_2}$ behaves almost like a standard Gaussian vector. We shall prove that, under the condition of the propositions, the term $\frac{\theta \theta^T}{\sigma^2 + \|\theta\|_2^2}$ in the covariance turns out to be negligible, whereas $\theta \|Y\|_2 / [\sigma^2 + \|\theta\|_2^2]$ is closely related to $\sqrt{n} \theta^* / \sigma$. The following lemma states that the conditional expectations of Z_f and $V(r_l, \omega_l)$ are almost the same as if the conditional covariance of $W / \|Y\|_2$ was the identity matrix. Recall the function g introduced in Section 2.2.3. Define the function $\Psi_l(x)$ by $\Psi_l(x) = \mathbb{E}[\eta_{r_l, \omega_l}(X)]$ where $X \sim \mathcal{N}(x, 1)$. As explained in Section 2.2.4 and proved in [17] (Section C.2.3), $\Psi_l(x) = \frac{1}{1 - 2\Phi(r_l)} \int_{-r_l}^{r_l} \phi(\xi) \cos(\xi x \frac{\omega_l}{r_l}) d\xi$. Obviously, we have $\Psi_l(0) = 1$. Besides it is also shown in [17] (Section C.2.3) that $-\frac{l}{k_0} \leq \Psi_l(x) \leq \frac{l}{k_0} + 2 \exp\left(-\frac{\omega_l^2 x^2}{r_l^2}\right)$.

Lemma 6. *If $s\|\theta\|_\infty \leq \sigma$, we have*

$$\left| \mathbb{E}_{\theta^*, \sigma}^{(3)} [Z_f | (\|Y\|_2, \theta)] - \sum_{i=1}^p \left[\mathbf{1}_{(\bar{\theta}_I)_i=0} g(sv_i) + \mathbf{1}_{(\bar{\theta}_I)_i \neq 0} \right] \right| \leq \frac{s^2}{5}. \quad (40)$$

Consider any $l \in \mathcal{L}_0$. If $\omega_l \|\theta\|_\infty \leq \sigma$, we have

$$\left| \mathbb{E}_{\theta^*, \sigma}^{(3)} [V(r_l, \omega_l) | (\|Y\|_2, \theta)] - \sum_{i=1}^p \left[\mathbf{1}_{(\bar{\theta}_I)_i=0} (1 - \Psi_l(v_j)) + \mathbf{1}_{(\bar{\theta}_I)_i \neq 0} \right] \right| \leq \frac{e^{1/2}}{2} \omega_l^2. \quad (41)$$

Also, the next lemma enforces that the deviations of the statistics Z_f and $V(r_l, \omega_l)$ are almost the same as if the conditional covariance of $W/\|Y\|_2$ was the identity matrix.

Lemma 7. *Assume that $\|\theta\|_\infty \leq [\sigma^2 + \|\theta\|_2^2]^{1/2}/s$. For any $t > 0$, one has*

$$\begin{aligned} \mathbb{P}_{\theta^*, \sigma}^{(3)} [Z_f - \mathbb{E}_{\theta^*, \sigma}^{(3)} [Z_f | (\|Y\|_2, \theta)] \geq se^{s^2/2} \sqrt{2pt} | (\|Y\|_2, \theta)] &\leq e^{-t}; \\ \mathbb{P}_{\theta^*, \sigma}^{(3)} [Z_f - \mathbb{E}_{\theta^*, \sigma}^{(3)} [Z_f | (\|Y\|_2, \theta)] \leq -se^{s^2/2} \sqrt{2pt} | (\|Y\|_2, \theta)] &\leq e^{-t}. \end{aligned} \quad (42)$$

Besides, for any $l \in \mathcal{L}_0$ and any $t > 0$, one has

$$\begin{aligned} \mathbb{P}_{\theta^*, \sigma}^{(3)} \left[V(r_l, \omega_l) - \mathbb{E}_{\theta^*, \sigma}^{(3)} [V(r_l, \omega_l) | (\|Y\|_2, \theta)] \geq \sqrt{2lp^{1/2}t} | (\|Y\|_2, \theta) \right] &\leq e^{-t}; \\ \mathbb{P}_{\theta^*, \sigma}^{(3)} \left[V(r_l, \omega_l) - \mathbb{E}_{\theta^*, \sigma}^{(3)} [V(r_l, \omega_l) | (\|Y\|_2, \theta)] \leq -\sqrt{2lp^{1/2}t} | (\|Y\|_2, \theta) \right] &\leq e^{-t}. \end{aligned} \quad (43)$$

Analysis of the tests under the null hypothesis. The assumptions of Lemma 2 are fulfilled. As a consequence, we have $\|\theta^* - \tilde{\theta}_I\|_\infty \leq \underline{c}^{(t)} \sigma \sqrt{\log(2p/\alpha)/n}$ with probability larger than $1 - \delta - \alpha/2$. Henceforth, we call this event \mathcal{B} and work conditionally to it. Thus, the support of $\tilde{\theta}_I$ is included in that of θ^* which in turn implies that $\sum_{j=1}^p \mathbf{1}_{\theta_j \neq 0} \mathbf{1}_{(\bar{\theta}_I)_j=0} + \mathbf{1}_{(\bar{\theta}_I)_j \neq 0} \leq k_0$ which implies $\|\theta\|_0 \leq k_0$.

Besides, we also have $\|\theta\|_\infty \leq \underline{c}^{(t)} \sigma \sqrt{\frac{\log(2p/\alpha)}{n}}$.

Since $\max_{l \in \mathcal{L}_0} \omega_l \leq s = \sqrt{\log(ek_0/\sqrt{p})} \vee 1 \leq (\underline{c}^{(t)})^{-1} \sqrt{\frac{n}{\log(2p/\alpha)}}$ for $n \geq 9\underline{c}^{(t)2} \log^2(ep/\alpha)$, it also follows from Assumption $\mathbf{A}[\alpha \wedge \delta]$ that

$$\frac{\|\theta\|_\infty}{\sigma} \leq \frac{1}{s} \wedge \left(\min_{l \in \mathcal{L}_0} \frac{1}{\omega_l} \right). \quad (44)$$

Thus, we are in position to apply Lemma 6. As explained in Section 2.2.3, we have $g(0) = 0$ and $g(x) \in [0, 1]$, it follows from that Lemma that $\mathbb{E}_{\theta^*, \sigma}^{(3)} [Z_f | (\|Y\|_2, \theta)] \leq k_0 + s^2/5$. Also since $1 - \Psi_l(0) = 0$ and $1 - \Psi_l(x) \in [0, 1 + \frac{l}{k_0}]$ we have

$$\mathbb{E}_{\theta^*, \sigma}^{(3)} [V[r_l, \omega_l] | (\|Y\|_2, \theta)] \leq k_0 + l + \frac{e^{1/2}}{2} \omega_l^2.$$

Then, we apply the deviation inequalities of Lemma 7 and integrate them with respect to $\|Y\|_2$ to conclude that

$$\mathbb{P}_{\theta^*, \sigma}^{(3)} [Z_f \geq k_0 + v_\alpha^{(f)} | \theta] \leq \alpha/2;$$

$$\sum_{l \in \mathcal{L}_0} \mathbb{P}_{\theta^*, \sigma}^{(3)} [V[r_l, \omega_l] \geq k_0 + l + v_{\alpha, l}^{(i)} | \theta] \leq \alpha/2 .$$

Taking the probability of the event \mathcal{B} into account, we conclude that the type I error probability of both tests is bounded by $\alpha + \delta$.

Analysis of the tests under the alternative hypothesis. Since $\|\theta^*\|_0$ is not too large, the assumptions of Lemma 2 are fulfilled. As under the null hypothesis, we have $\|\theta^* - \tilde{\theta}_{\mathbf{I}}\|_\infty \leq \underline{c}^{(t)} \sigma \sqrt{\log(2p/\alpha)/n}$ with probability higher than $1 - \delta - \alpha/2$ and we still work conditionally to this event called \mathcal{B} . If $(\tilde{\theta}_{\mathbf{I}})_{(k_0+1)} \geq \underline{c}^{(t)} \sigma \sqrt{\log(2p/\alpha)/n}$, then both tests reject the null hypothesis, so that we can assume henceforth that $(\tilde{\theta}_{\mathbf{I}})_{k_0+1} = 0$.

Since (44) is still valid, we are in position to apply again Lemmas 6 and 7. Hence, conditionally on θ and $\|Y\|_2$, we have

$$Z_f \geq \|\bar{\theta}_{\mathbf{I}}\|_0 + \sum_{i=1}^p g(sv_i) \mathbf{1}_{(\bar{\theta}_{\mathbf{I}})_i=0} - \frac{s^2}{5} - se^{s^2/2} \sqrt{2p \log\left(\frac{2}{\alpha}\right)},$$

with probability higher than $1 - \alpha/2$. Define \tilde{v} by

$$\tilde{v}_i = +\infty \text{ if } (\bar{\theta}_{\mathbf{I}})_i \neq 0 \quad \text{and} \quad \tilde{v}_i = v_i \text{ if } (\bar{\theta}_{\mathbf{I}})_i = 0 .$$

Recall that $\lim_{x \rightarrow +\infty} g(x) = 1$ and $\forall x \in \mathbb{R}, 0 \leq g(x) \leq 1$ (see [17]). So it holds that

$$Z_f \geq \sum_{i=1}^p g(s\tilde{v}_i) - \frac{s^2}{5} - se^{s^2/2} \sqrt{2p \log\left(\frac{2}{\alpha}\right)}. \quad (45)$$

Also, for any $l \in \mathcal{L}_0$, we have

$$V[r_l, \omega_l] \geq \|\bar{\theta}_{\mathbf{I}}\|_0 + \sum_{i=1}^p (1 - \Psi_l(\tilde{v}_i)) \mathbf{1}_{(\bar{\theta}_{\mathbf{I}})_i=0} - \frac{\sqrt{e}}{2} \omega_l^2 - \sqrt{2p^{1/2} l \log\left(\frac{2}{\alpha}\right)},$$

with probability larger than $1 - \alpha/2$. As above, we have $\lim_{x \rightarrow \infty} \Psi_l(x) = 0$, and so

$$V[r_l, \omega_l] \geq \sum_{i=1}^p (1 - \Psi_l(\tilde{v}_i)) - \frac{\sqrt{e}}{2} \omega_l^2 - \sqrt{2p^{1/2} l \log\left(\frac{2}{\alpha}\right)}, \quad (46)$$

In the sequel, we show that (45) and (46) imply the desired type II error probability bounds.

Case 1: Analysis of (45) for $\phi^{(f)}$. Write $\bar{s} = \sqrt{\log(e \frac{k_0^2}{p})} \vee 1$ the tuning parameter used in [17] for the corresponding test in the Gaussian sequence model. Note that $s \geq \bar{s}/\sqrt{2}$, $se^{s^2/2} \leq 2\bar{s}^{-1} e^{\bar{s}^2/2}$, and $s^2/5 \leq se^{s^2/2}$. We have shown in the proof of Proposition 2 in [17] that for a vector $x \in \mathbb{R}^p$ and any $\alpha \in (0, 1)$

$$\sum_{i=1}^p g(\bar{s}x_i) \geq k_0 + \bar{s}^{-1} e^{\bar{s}^2/2} 2\sqrt{8p \log(2/\alpha)},$$

as soon as one of the two following condition holds for constants $c_\alpha, c'_\alpha, c''_\alpha$ positive and large enough, depending only on α

$$|x_{(k_0+q)}| \geq c_\alpha \sqrt{\frac{k_0}{q \log(1 + k_0/\sqrt{p})}}, \text{ for some } q \geq c'_\alpha \frac{k_0}{\sqrt{\log(1 + \frac{k_0^2}{p})}} ;$$

$$\sum_{i=1}^p [x_i^2 \wedge \bar{s}^{-1}] \geq c''_\alpha \frac{1}{\log(1 + k_0/\sqrt{p})} .$$

It then follows from (45), that, given θ and $\|Y\|_2^2$ satisfying \mathcal{B} , the test rejects the null with probability higher than $1 - \alpha/2$ if

$$|\tilde{v}_{(k_0+q)}| \geq c_\alpha \sqrt{\frac{k_0}{q \log(1 + k_0/\sqrt{p})}}, \text{ for some } q \geq c'_\alpha \frac{k_0}{\sqrt{\log(1 + \frac{k_0^2}{p})}}; \quad (47)$$

$$\sum_{i=1}^p [\tilde{v}_i^2 \wedge \bar{s}^{-1}] \geq c''_\alpha \frac{1}{\log(1 + k_0/\sqrt{p})}. \quad (48)$$

Recall that, for $i \notin \mathcal{S}(\bar{\theta}_\mathbf{I})$, $\tilde{v}_i = v_i = \theta_i \|Y\|_2 / (\sigma^2 + \|\theta\|_2^2) = \theta_i^* \|Y\|_2 / (\sigma^2 + \|\theta\|_2^2)$. Since $\|Y\|_2^2 / (\sigma^2 + \|\theta\|_2^2)$ follows a χ^2 distribution with $n/3$ degrees of freedom, we have $\|Y\|_2^2 \geq n(\sigma^2 + \|\theta\|_2^2)/6$ with probability higher than $1 - e^{-n/27}$ (see Lemma 3). This implies that for any $i = 1, \dots, p$, we have

$$|\tilde{v}_i| \geq |\theta_i^*| \sqrt{\frac{n}{6(\sigma^2 + \|\theta\|_2^2)}}.$$

Lemma 8. *Assume that the event \mathcal{B} holds, that $d_2^2(\theta^*; \mathbb{B}_0[k_0]) \leq \sigma^2$, and that $(\bar{\theta}_\mathbf{I})_{k_0+1} = 0$. We have $\|\theta\|_2^2 \leq c'\sigma^2$.*

As a consequence, on the intersection of \mathcal{B} and an event of probability higher than $1 - e^{-n/27}$, we have

$$|\tilde{v}_i| \geq c \frac{\sqrt{n}}{\sigma} |\theta_i^*|.$$

Together with (47) and (48), we have characterized the type II error probability of $\phi^{(f)}$.

Case 2: Analysis of (46) for $\phi^{(i)}$. Observe $\sqrt{e}\omega_l^2/2$ is at most of the order of $\log(p)$ and is therefore negligible compared to $\sqrt{p^{1/2}l}$. We have shown in the proof of Proposition 3 in [17] that, for a vector $x \in \mathbb{R}^p$, and for any α in $(0, 1)$ we have

$$\sum_{i=1}^n \Psi_l(x_i) \geq k_0 + 2l + 2\sqrt{2lp^{1/2}} \sqrt{\log\left(\frac{\pi^2[1 + \log_2(l/l_0)]}{3\alpha}\right)},$$

for some $l \in \mathcal{L}_0$, if for constants c_α, c'_α positive and large enough, depending only on α

$$|x_{(k_0+q)}| \geq c_\alpha \frac{1 + \log\left(\frac{k_0}{q \wedge k_0}\right)}{\sqrt{\log\left(1 + \frac{k_0}{\sqrt{p}}\right)}}, \text{ for some } q \geq c'_\alpha k_0^{4/5} p^{1/10}. \quad (49)$$

Actually, in Proposition 3 in [17], we had considered a wider range of q 's as the collection \mathcal{L}_0 was slightly larger, but this does not change the arguments here. In our setting, Condition (49) and (46) imply that $V[r_l, \omega_l] \geq k_0 + l + v_{\alpha, l}^{(i)}$ for some $l \in \mathcal{L}_0$ if

$$|\tilde{v}_{(k_0+q)}| \geq c_\alpha \frac{1 + \log\left(\frac{k_0}{q \wedge k_0}\right)}{\sqrt{\log\left(1 + \frac{k_0}{\sqrt{p}}\right)}}, \text{ for some } q \geq c'_\alpha k_0^{4/5} p^{1/10}. \quad (50)$$

Then, arguing as in Case 1, we have $|\tilde{v}_i| \geq c'|\theta_i|/\sigma$ on the intersection of \mathcal{B} and an event of probability higher than $1 - e^{-n/27}$. Putting everything together, we have controlled the type II error probability of $\phi^{(i)}$.

Proof of Lemma 6. In view of the conditional distribution of W_j given Y , one has

$$\begin{aligned}\mathbb{E}_{\theta^*, \sigma}^{(3)}[\varphi(s; W_j / \|\bar{Y}^{(3)}\|_2) | (\|Y\|_2, \theta)] &= \int_{-1}^1 (1 - |\xi|) \cos \left[\xi \frac{s\theta_j \|Y\|_2}{\sigma^2 + \|\theta\|_2^2} \right] \exp \left[\xi^2 \frac{s^2 \theta_j^2}{2(\sigma^2 + \|\theta\|_2^2)} \right] d\xi \\ &= g(sv_j) + \int_{-1}^1 (1 - |\xi|) \cos \left[\xi \frac{s\theta_j \|Y\|_2}{\sigma^2 + \|\theta\|_2^2} \right] \left(\exp \left[\xi^2 \frac{s^2 \theta_j^2}{2(\sigma^2 + \|\theta\|_2^2)} \right] - 1 \right) d\xi.\end{aligned}$$

Since $s\|\theta\|_\infty \leq \sigma$, the remainder term is (in absolute value) less than

$$2 \int_0^1 (1 - \xi) \left(e^{\frac{\xi s^2 \theta_j^2}{2\sigma^2}} - 1 \right) d\xi \leq 2e^{1/2} \int_0^1 (1 - \xi) \xi^2 \frac{s^2 \theta_j^2}{2(\sigma^2 + \|\theta\|_2^2)} d\xi \leq \frac{s^2 \theta_j^2}{5(\sigma^2 + \|\theta\|_2^2)}.$$

Summing over all $j = 1, \dots, p$ such that $(\bar{\theta}_{\mathbf{I}})_j = 0$, we obtain the first result of Lemma 6. Turning to $V[r_l, \omega_l]$, we have

$$\mathbb{E}_{\theta^*, \sigma}^{(3)} \left[\eta_{r_l, \omega_l} \left(\frac{W_j}{\|Y\|_2} \right) | (\|Y\|_2, \theta) \right] = \frac{r_l}{1 - 2\bar{\Phi}(r_l)} \int_{-1}^1 \frac{e^{-r_l^2 \xi^2 / 2}}{\sqrt{2\pi}} \cos \left(\xi \omega_l \frac{\theta_j \|Y\|_2}{\sigma^2 + \|\theta\|_2^2} \right) \exp \left[\xi^2 \frac{\omega_l^2 \theta_j^2}{2(\sigma^2 + \|\theta\|_2^2)} \right] d\xi$$

As a consequence,

$$\begin{aligned}\left| \mathbb{E}_{\theta^*, \sigma}^{(3)} \left[\eta_{r_l, \omega_l} \left(\frac{W_j}{\|Y\|_2} \right) | (\|Y\|_2, \theta) \right] - \Psi_l(v_j) \right| &\leq \frac{r_l}{1 - 2\bar{\Phi}(r_l)} \int_{-1}^1 \frac{e^{-r_l^2 \xi^2 / 2}}{\sqrt{2\pi}} \left| e^{\xi^2 \omega_l^2 \frac{\theta_j^2}{2(\sigma^2 + \|\theta\|_2^2)}} - 1 \right| d\xi \\ &\leq \frac{e^{1/2} r_l}{1 - 2\bar{\Phi}(r_l)} \int_{-1}^1 \frac{e^{-r_l^2 \xi^2 / 2}}{\sqrt{2\pi}} \xi^2 \omega_l^2 \frac{\theta_j^2}{2(\sigma^2 + \|\theta\|_2^2)} d\xi \\ &\leq \frac{e^{1/2} \omega_l^2 \theta_j^2}{2(\sigma^2 + \|\theta\|_2^2)},\end{aligned}$$

where we used the condition $\omega_l \|\theta\|_\infty \leq \sigma$ in the second line. Summing this bound over all j such that $(\bar{\theta}_{\mathbf{I}})_j = 0$ yields the desired result. \square

Proof of Lemma 7. We shall apply the Gaussian concentration theorem (see e.g. [5]) to both Z_f and $V[r_l, \omega_l]$. The covariance matrix $\mathbf{\Gamma}$ associated to the conditional distribution of $W/\|Y\|_2$ decomposes as $\mathbf{I}_p - a \frac{\theta}{\|\theta\|_2} \frac{\theta^T}{\|\theta\|_2}$ with $a = \|\theta\|_2^2 / [\sigma^2 + \|\theta\|_2^2] \in [0, 1)$ and in particular its operators norm is less than one. Write $\mathbf{\Gamma}^{1/2}$ for a square-root of this matrix and let U denote a standard Gaussian vector. Conditionally to Y , $W/\|Y\|_2$ is distributed as $v + \mathbf{\Gamma}^{1/2}U$. For any $u \in \mathbb{R}^p$, define

$$\zeta(u) = \sum_{j=1}^p \mathbf{1}_{(\bar{\theta}_{\mathbf{I}})_j = 0} \int_{-1}^{+1} (1 - |\xi|) \cos(\xi s[\mu_j + (\mathbf{\Gamma}^{1/2}u)_j]) e^{\xi^2 s^2 / 2} d\xi.$$

Given two vectors u and u' , one has

$$\begin{aligned}|\zeta(u) - \zeta(u')| &\leq \sum_{j=1}^p \mathbf{1}_{(\bar{\theta}_{\mathbf{I}})_j = 0} \int_{-1}^{+1} (1 - |\xi|) \left| \cos(\xi s[\mu_j + (\mathbf{\Gamma}^{1/2}u)_j]) - \cos(\xi s[\mu_j + (\mathbf{\Gamma}^{1/2}u')_j]) \right| e^{\xi^2 s^2 / 2} d\xi \\ &\leq s e^{s^2 / 2} \sum_{j=1}^p |\mathbf{\Gamma}^{1/2}(u - u')_j| = s e^{s^2 / 2} \|\mathbf{\Gamma}^{1/2}(u - u')\|_1 \\ &\leq s e^{s^2 / 2} \sqrt{p} \|u - u'\|_2,\end{aligned}$$

since the cosine function is 1-Lipschitz. As a consequence, the function $u \mapsto Z(u)$ is $se^{s^2/2}\sqrt{p}$ -Lipschitz. The deviation inequalities (42) then follow from the Gaussian concentration theorem (see e.g. [5]).

As for $V[r_l, \omega_l]$, we argue similarly that, for $\omega_l > r_l$, it is conditionally distributed as a Lipschitz function of a standard Gaussian vector with Lipschitz constant

$$\frac{r_l \omega_l}{1 - 2\overline{\Phi}(r_l)} \int_{-1}^1 e^{\xi^2(\omega_l^2 - r_l^2)/2} \frac{\sqrt{p}}{\sqrt{2\pi}} d\xi \leq \sqrt{\frac{8p}{\pi}} \frac{r_l \omega_l}{(1 - 2\overline{\Phi}(r_l))(\omega_l^2 - r_l^2)} e^{(\omega_l^2 - r_l^2)/2}.$$

Since $l \geq k_0^{4/5} p^{1/10}$, we have $\omega_l^2 - r_l^2 \geq 2\omega_l$ for any $l \in \mathcal{L}_0$ and the Lipschitz constant is therefore less than

$$\frac{r_l}{(1 - 2\overline{\Phi}(r_l))} \sqrt{\frac{2p}{\pi}} e^{(\omega_l^2 - r_l^2)/2} \leq p^{1/4} l^{1/2},$$

where the last inequality is a consequence of the definition of r_l and ω_l and is detailed in the proof of Lemma 6 in [17]. □

Proof of Lemma 8. Under \mathcal{B} , we have $\|\theta^* - \tilde{\theta}_{\mathbf{I}}\|_{\infty} \leq \underline{c}^{(t)} \sigma \sqrt{\log(2p/\alpha)/n}$. Hence,

$$\begin{aligned} \|\theta\|_2^2 &= \sum_{i: \bar{\theta}_{\mathbf{I}} \neq 0} (\theta_i^* - (\bar{\theta}_{\mathbf{I}})_i)^2 + \sum_{i: \bar{\theta}_{\mathbf{I}} = 0} \theta_i^{*2} \\ &\leq k_0 \underline{c}^{(t)2} \sigma^2 \frac{\log(2p/\alpha)}{n} + \sum \theta_i^{*2} \mathbf{1}_{|\theta_i^*| \leq 2\underline{c}^{(t)} \sigma \sqrt{\log(2p/\alpha)/n}} \\ &\leq 5k_0 \underline{c}^{(t)2} \sigma^2 \frac{\log(2p/\alpha)}{n} + \sigma^2 \leq c' \sigma^2, \end{aligned}$$

where we used in the second line the definition of $\bar{\theta}_{\mathbf{I}}$ and $(\bar{\theta}_{\mathbf{I}})_{k_0+1} = 0$ and we used $d_2^2(\theta^*; \mathbb{B}_0[k_0]) \leq \sigma^2$ together with Assumption $(\mathbf{A}[\alpha \wedge \delta])$ in the last line. □

5.4 Proof of Theorem 2

Consider any $\theta \in \mathbb{B}_0[k_0]$. In view of Propositions 3–5, we can bound the rejection probability as follows

$$\mathbb{P}_{\theta, \sigma}[\phi^{(ag)} = 1] \leq 4(\delta + \alpha) + \mathbb{P}_{\theta, \sigma}[d_2^2(\hat{\theta}_{SL}; \mathbb{B}_0[k_0]) \geq \sigma^2/2] \quad (51)$$

Since, under the null hypothesis, θ^* is k_0 -sparse, we have

$$d_2^2(\hat{\theta}_{SL}; \mathbb{B}_0[k_0]) \leq \|\theta^* - \hat{\theta}_{SL}\|_2^2.$$

Applying Lemma 1, we derive that, with probability higher than $1 - \delta$, $d_2^2(\hat{\theta}_{SL}; \mathbb{B}_0[k_0]) \leq \underline{c}_1^{SL} \sigma^2 \frac{k_0}{n} \log(p/\delta)$. Thus, by Condition $\mathbf{A}[\alpha \wedge \delta]$, we have $\mathbb{P}_{\theta, \sigma}[d_2^2(\hat{\theta}_{SL}; \mathbb{B}_0[k_0]) \geq \sigma^2/2] \leq \delta$. From (51), we derive that $\mathbb{P}_{\theta, \sigma}[\phi^{(ag)} = 1] \leq 5\delta + 4\alpha$. Looking more closely at the proof of Propositions 3–5, we observe that each occurrence of the probability δ corresponds to the same control of the square-root Lasso estimator $\hat{\theta}_{SL}$. As a consequence $\phi^{(ag)}$ satisfies $(\mathbf{P}_1[\delta + 4\alpha])$. Turning to the Type II error, we fix $\Delta \leq p - k_0$ and assume that $\theta^* \in \mathbb{B}_0[k_0 + \Delta]$.

Case 1: $\Delta \leq cn/\log(p/\delta)$. If $k_0 \leq p^{1/2-\varsigma}$, then the squared separation distance $\min[\Delta \log(p)/n, 1/\sqrt{n} + k_0 \log(p)/n]$ in (20) is a consequence of Propositions 2 and 3 and is achieved by the combination of $\phi^{(t)}$ and $\phi^{(\chi)}$. If $k_0 \geq p^{1/2+\varsigma}$, the squared separation distance $\Delta \log(p)/n$ is still achieved by

$\phi^{(t)}$. To prove the last part of the result, let us assume that θ^* is such that $\max(\phi^{(t)}, \phi^{(x)}, \phi^{(f)}, \phi^{(i)})$ does not reject the null with high probability. We shall prove that this implies $d_2^2(\theta^*; \mathbb{B}_0[k_0]) \leq c_{\alpha, \varsigma} \sigma^2 k_0 / [\log(p)n]$. From Proposition 3, we have $d_2^2(\theta^*; \mathbb{B}_0[k_0]) \leq \sigma^2$. In view of Proposition 4, we have

$$\sum_i \theta_i^{*2} \wedge \frac{\sigma^2}{n \log(p)} \leq c_{\alpha, \varsigma} \sigma^2 \frac{k_0}{n \log(p)}.$$

In view of Proposition 5, we have

$$|\theta_{(k_0+q)}^*| \leq c_{\alpha, \varsigma} \sigma \frac{1 + \log(\frac{k_0}{q \wedge k_0})}{\sqrt{n \log(p)}},$$

for all $q \geq c'_\alpha k_0^{4/5} p^{1/10}$. Finally, Proposition 2 enforces that

$$|\theta_{(k_0+q)}^*| \leq c_{\alpha, \varsigma} \sigma \sqrt{\frac{\log(p)}{n}},$$

for all $q < c'_\alpha k_0^{4/5} p^{1/10}$. Putting everything together, we obtain

$$\begin{aligned} d_2^2(\theta^*; \mathbb{B}_0[k_0]) &\leq \sum_{q=1}^{\Delta} \theta_{(k_0+q)}^{*2} \leq c_{\alpha, \varsigma} \sigma^2 \left[\frac{k_0}{n \log(p)} + \sum_{q=k_0/\log^2(p)}^{k_0} \frac{1 + \log^2(\frac{k_0}{q \wedge k_0})}{n \log(p)} + \sum_{q=k_0}^{\Delta} \theta_i^{*2} \right] \\ &\leq c'_{\alpha, \varsigma} \sigma^2 \frac{k_0}{n \log(p)}, \end{aligned}$$

where we used that, for $q \geq k_0$, $|\theta_{(k_0+q)}^*|$ is small compared to $\sigma/\sqrt{n \log(p)}$. This concludes the proof for Case 1.

Case 2: $\Delta \geq cn/\log(p/\delta)$. In that case, $\Delta \log(p)/n$ is larger than $k_0 \log(p)/n + n^{-1/2}$ and the first result in (20) is a consequence of the analysis of $\phi^{(x)}$ in Proposition 3. We now turn to the case $k_0 \geq p^{1/2+\varsigma}$ and we need to prove that the squared separation distance is less than $c_{\alpha, \delta} \sigma^2 k_0 / [n \log(p)]$. If $d_2^2(\theta^*; \mathbb{B}_0[k_0]) \geq \sigma^2$, then $\phi^{(x)}$ rejects the null hypothesis with high probability. Thus, we can assume that $d_2^2(\theta^*; \mathbb{B}_0[k_0]) \leq \sigma^2$. Also, we can assume that $\|\widehat{\theta}_{SL} - \widetilde{\theta}_{SL, k_0}\|_2^2 \leq \sigma^2/2$, otherwise the test $\phi^{(ag)}$ rejects the null. Finally, we can assume that $\|\theta^* - \widetilde{\theta}_{SL, k_0}\|_2^2 \leq \sigma^2/2$, otherwise the test $\phi^{(x)}$ also rejects the null with high probability. By triangular inequality, θ^* therefore satisfies $\|\theta^* - \widehat{\theta}_{SL}\|_2^2 \leq 2\sigma^2$ and we are in position to apply Lemma 2, which implies

$$\|\theta^* - \widetilde{\theta}_{\mathbf{I}}\|_\infty \leq c_1 \sigma \sqrt{\frac{\log(2p/\alpha)}{n}}, \quad (52)$$

with probability higher than $1 - \alpha/2$ conditionally to $\widehat{\theta}_{SL}$. As a consequence, the event \mathcal{B} involved in the proof of Propositions 4 and 5 is true. As ensuring this event is the only occurrence in the proof of these propositions where the restrictions $\|\theta^*\|_0 \leq cn/\log(p/\delta)$ is needed, we conclude that, given \mathcal{B} , $\max(\phi^{(f)}, \phi^{(i)})$ rejects the null with probability higher than $1 - \alpha/2$ if any of the conditions (14), (15), or (19) is satisfied. Similarly, Condition (52) (with $\alpha/2$ replaced by α) allows to adapt the proof of Proposition 2 without the restriction on $\|\theta^*\|_0$. Thus, $\phi^{(t)}$ rejects the null with conditional probability higher than $1 - \alpha$ under (10).

Arguing as Case 1, we conclude that the aggregated test rejects the null with high probability if $d_2^2(\theta^*; \mathbb{B}_0[k_0])$ is large compared to $\sigma^2 k_0 / [n \log(p)]$.

5.5 Proof of Theorem 3

Let d denote any positive integer. Let $S \subset \{1, \dots, d\}$. For $u \in \mathbb{R}^d$, we write $u_S = (u_i \mathbf{1}_{i \in S})_i$ for the vector in \mathbb{R}^d whose values outside S have been set to 0.

These notation are also extended to matrices. Given r a positive integer and a $r \times d$ matrix \mathbf{M} , we write \mathbf{M}_S for the $r \times d$ matrix defined by $(\mathbf{M}_S)_{i \leq r, j \leq d} = (\mathbf{M}_{i,j} \mathbf{1}_{\{j \in S\}})_{i \leq r, j \leq d}$. For $R \subset \{1, \dots, r\}$, we also write $\mathbf{M}_{R,S}$ for the $r \times d$ -dimensional matrix such that $(\mathbf{M}_{R,S})_{i \leq r, j \leq d} = (\mathbf{M}_{i,j} \mathbf{1}_{\{i \in R, j \in S\}})_{i \leq r, j \leq d}$.

5.5.1 Proof of Theorem 3

Let $\delta > 0$ and consider any subset S satisfying the property $(\mathbf{S}[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3])$.

Lemma 9. *The exists a constant c such that the following holds for all $\delta > 0$. If*

$$c \left[\mathbf{a}_2 \|\theta^*\|_0 + \log \left(\frac{4}{\delta} \right) \right] \leq m ,$$

there exists an event \mathcal{B}_1 of probability higher than $1 - \delta/2$ such that

$$(2\eta)^{-1} \leq \frac{1}{m} \lambda_{\min,S}(\mathbf{X}_S^{(0)T} \mathbf{X}_S^{(0)}) \leq \frac{1}{m} \lambda_{\max,S}(\mathbf{X}_S^{(0)T} \mathbf{X}_S^{(0)}) \leq 2\eta ; \quad (53)$$

$$\left\| \frac{1}{m} \mathbf{X}_S^{(0)T} \mathbf{X}_S^{(0)} \theta_S^* \right\|_2 \leq 2\eta \sqrt{\log(4/\delta)} \|\theta_S^*\|_2 , \quad (54)$$

where $\lambda_{\min,S}$ and $\lambda_{\max,S}$ respectively refer to the smallest and largest eigenvalue of a matrix restricted to its coordinates in $S \times S$.

So on the event \mathcal{B}_1 defined above, the matrix $\mathbf{X}_S^{(0)T} \mathbf{X}_S^{(0)}$ restricted to its coordinates in $S \times S$ is non-singular. Recall that the matrix $\mathbf{X}_S^{(0)T} \mathbf{X}_S^{(0)}$ is 0 outside $S \times S$. Nevertheless, we can define its pseudo-inverse $(\mathbf{X}_S^{(0)T} \mathbf{X}_S^{(0)})^{-1}$ by considering its inverse when restricted to $S \times S$ and fixing all its remaining entries to 0. The restricted least-squares estimator $\hat{\theta}_{l_s,S}$ is then conditionally distributed as follows

$$(\hat{\theta}_{l_s,S} | \mathbf{X}^{(0)}) \sim \mathcal{N}(\theta_S^* + (\mathbf{X}_S^{(0)T} \mathbf{X}_S^{(0)})^{-1} \mathbf{X}_S^{(0)T} \mathbf{X}_S^{(0)} \theta_S^*, \sigma^2 (\mathbf{X}_S^{(0)T} \mathbf{X}_S^{(0)})^{-1}). \quad (55)$$

Define the bias $B = \theta^* - \mathbb{E}^{(0)}[\hat{\theta}_{l_s,S} | \mathbf{X}^{(0)}]$. On the event \mathcal{B}_1 , it follows from the definition in Equation (25) of $\mathbf{S}[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$ and Lemma 9 that

$$\|B\|_2^2 \leq 16\eta^4 \log(4/\delta) \|\theta_S^*\|_2^2 + \|\theta_S^*\|_2^2 \leq 17\eta^4 \log(4/\delta) \|\theta_S^*\|_2^2 \leq 17\eta^4 \log(4/\delta) \mathbf{a}_3^2 M(\mathbf{a}_1, \frac{\theta^*}{\sigma}) \frac{\log(p)}{m} \sigma^2 . \quad (56)$$

Next, since $\hat{\theta}_{l_s,S}$ follows a normal distribution (55), we can easily bound its deviations. In particular, we deduce from (53) that there exists an event \mathcal{B}_2 of probability higher than $1 - \delta/3$ such that on $\mathcal{B}_1 \cap \mathcal{B}_2$, one has

$$|(\hat{\theta}_{l_s,S})_i - \theta_i^*| \leq \sigma \sqrt{2\eta \frac{\log(6p/\delta)}{m}} + |B_i|, \quad \text{for } i \in S \quad \text{and} \quad |(\hat{\theta}_{l_s,S})_i - \theta_i^*| = |\theta_i^*| = |B_i| \quad \text{otherwise} . \quad (57)$$

Lemma 10. *Assume that $\log(6/\delta) \leq cn$. There exists an event \mathcal{B}_3 of probability higher than $1 - \delta/6$ such that on $\mathcal{B} = \cap_{i=1}^3 \mathcal{B}_i$, we have*

$$\frac{\sigma^2}{2} \leq \hat{\sigma}_S^2 \leq 2\sigma^2 \left[1 + \eta \mathbf{a}_3^2 M(\mathbf{a}_1, \frac{\theta^*}{\sigma}) \frac{\log(p)}{m} \right]$$

Putting everything together, we derive that, under \mathcal{B} , one has

$$\frac{|(\widehat{\theta}_{l_s, S})_i - \theta_i^*|}{\widehat{\sigma}_S} \leq 2\sqrt{\eta \frac{\log(6p/\delta)}{m}} + \sqrt{2} \frac{|B_i|}{\sigma}, \quad i = 1, \dots, p.$$

This implies that, for all $i = 1, \dots, p$,

$$\frac{|\theta_i^*|}{\sqrt{2}\sigma[1 + \eta\mathbf{a}_3^2 M(\mathbf{a}_1, \frac{\theta^*}{\sigma}) \frac{\log(p)}{m}]^{1/2}} - 2\sqrt{\eta \frac{\log(6p/\delta)}{m}} - \sqrt{2} \frac{|B_i|}{\sigma} \leq \frac{|(\widehat{\theta}_{l_s, S})_i|}{\widehat{\sigma}_S} \leq \sqrt{2} \frac{|\theta_i^*|}{\sigma} + 2\sqrt{\eta \frac{\log(6p/\delta)}{m}} + \sqrt{2} \frac{|B_i|}{\sigma}. \quad (58)$$

Under the null hypothesis. Suppose that $\theta^* \in \mathbb{B}_0[k_0]$. Note that (27) implies that

$$\underline{c}_* \geq \sqrt{2}\mathbf{a}_1 + t_1 + 2\sqrt{\eta \frac{\log(6p/\delta)}{\log(p)}} \text{ with } t_1 = 2\eta^2 \mathbf{a}_3 \sqrt{17 \log(\frac{4}{\delta})}.$$

Assume that θ^* belongs to $\mathbb{B}_0[k_0]$. From (58), we deduce that, conditionally on the event \mathcal{B} , one has

$$\begin{aligned} N[\underline{c}_*; \widehat{\theta}_{l_s, S}/\widehat{\sigma}_S] &\leq \left| \left\{ i : \frac{\sqrt{2}|\theta_i^*|}{\sigma} \geq \sqrt{2}\mathbf{a}_1 \sqrt{\frac{\log(p)}{m}} \right\} \right| + \left| \left\{ i : \frac{\sqrt{2}|B_i|}{\sigma} \geq t_1 \sqrt{\frac{\log(p)}{m}} \right\} \right| \\ &\leq k_0 - M[\mathbf{a}_1, \frac{\theta^*}{\sigma}] + \frac{2\|B\|_2^2 m}{\sigma^2 t_1^2 \log(p)} \\ &\stackrel{(56)}{\leq} k_0 - M[\mathbf{a}_1, \frac{\theta^*}{\sigma}] + \frac{34\eta^4}{t_1^2} \log(\frac{4}{\delta}) \mathbf{a}_3^2 M(\mathbf{a}_1, \frac{\theta^*}{\sigma}) \leq k_0. \end{aligned}$$

As a consequence, the test accepts the null hypothesis under the event \mathcal{B} .

Under the alternative hypothesis. We now assume that θ^* belongs to $\mathbb{B}_0[k_0 + \Delta]$ and satisfies

$$d_2^2(\theta^*; \mathbb{B}_0[k_0]) \geq \underline{t}^2 \sigma^2 \frac{\Delta \log(p)}{m} \quad \text{with} \quad \underline{t} = 4\sqrt{2} \left[1 + \eta\mathbf{a}_3^2 (k_0 + \Delta) \frac{\log(p)}{m} \right]^{1/2} \left[\underline{c}_* \vee \left(5\eta^2 \sqrt{\log(4/\delta)} \mathbf{a}_3 \right) \right]. \quad (59)$$

Consider the set $T = \{i, |\theta_i^*| \geq \sigma \underline{t} \sqrt{\frac{\log(p)}{2m}}\}$ of large coordinates of θ^* . In view of $d_2^2(\theta^*; \mathbb{B}_0[k_0])$, we have

$$d_2^2(\theta_T^*, \mathbb{B}_0[k_0]) \geq \underline{t}^2 \sigma^2 \frac{\Delta \log(p)}{2m}. \quad (60)$$

On the event \mathcal{B} , it follows from (58) and the definition of \underline{c}_* that $|(\widehat{\theta}_{l_s, S})_i|/\widehat{\sigma}_S \geq \underline{c}_* \sqrt{\log(p)/m}$ if

$$\frac{|\theta_i^*|}{\sigma} \geq 4\sqrt{2} [1 + \eta\mathbf{a}_3^2 M(\mathbf{a}_1, \frac{\theta^*}{\sigma}) \frac{\log(p)}{m}]^{1/2} \underline{c}_* \sqrt{\frac{\log(p)}{m}} \quad \text{and} \quad |B_i| \leq \frac{|\theta_i^*|}{4[1 + \eta\mathbf{a}_3^2 M(\mathbf{a}_1, \frac{\theta^*}{\sigma}) \frac{\log(p)}{m}]^{1/2}}.$$

Observe that $\underline{t} \geq 4\sqrt{2} [1 + \eta\mathbf{a}_3^2 M(\mathbf{a}_1, \frac{\theta^*}{\sigma}) \frac{\log(p)}{m}]^{1/2} \underline{c}_*$. Denoting $T_0 = T \cap \{i : 4|B_i| \geq |\theta_i^*| [1 + \eta\mathbf{a}_3^2 M(\mathbf{a}_1, \frac{\theta^*}{\sigma}) \frac{\log(p)}{m}]^{-1/2}\}$, we obtain that $N[\underline{c}_*; \widehat{\theta}_{l_s, S}/\widehat{\sigma}_S] \geq |T| - |T_0|$. We can bound $\|\theta_{T_0}^*\|_2^2$ in terms of the bias $\|B\|_2^2$ and then use (56) and (59).

$$\begin{aligned} \|\theta_{T_0}^*\|_2^2 &\leq 16 [1 + \eta\mathbf{a}_3^2 M(\mathbf{a}_1, \frac{\theta^*}{\sigma}) \frac{\log(p)}{m}] \|B\|_2^2 \\ &\leq 272\eta^4 \log(4/\delta) [1 + \eta\mathbf{a}_3^2 M(\mathbf{a}_1, \frac{\theta^*}{\sigma}) \frac{\log(p)}{m}] \mathbf{a}_3^2 \frac{M[\mathbf{a}_1, \frac{\theta^*}{\sigma}] \log(p)}{m} \sigma^2 < \underline{t}^2 \sigma^2 \frac{\Delta \log(p)}{2m}, \end{aligned}$$

where the inequality $M[\mathbf{a}_1, \frac{\theta^*}{\sigma}] \leq \Delta$ is a consequence of (59) and $\underline{c}_* \geq \sqrt{2}\mathbf{a}_1$. In view of Equation (60), we have $\|\theta_{T_0}^*\|_2^2 < d_2^2(\theta_T^*, \mathbb{B}_0[k_0])$, which implies $|T_0| < |T| - k_0$ and therefore $N[\underline{c}_*; \hat{\theta}_{t_s, S}/\hat{\sigma}] > k_0$. The test therefore rejects the null hypothesis under the event \mathcal{B} , which concludes the proof.

Proof of Lemma 9. We first show (53). Recall that $\mathbf{X}^{(0)}$ is independent of S and that the restriction of $\mathbf{N} = \Sigma_{S,S}^{-1/2} \mathbf{X}_S^{(0)T} \mathbf{X}_S^{(0)} \Sigma_{S,S}^{-1/2}$ to $S \times S$ follows a standard Wishart distribution - all coordinates outside $S \times S$ being 0. by $\mathbf{S}[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$, the size of the corresponding covariance matrix is less than $|S| \leq \mathbf{a}_2 \|\theta^*\|_0$. From e.g. [21], we deduce, on an event \mathcal{B}_{1-1} of probability larger than $1 - \delta/4$, we have

$$\begin{aligned} \eta^{-1} \left(1 - c_R \sqrt{\frac{|S| \log(4/\delta)}{m}} - c_R \frac{\log(4/\delta) + |S|}{m} \right) &\leq \frac{1}{m} \lambda_{\min, S}(\mathbf{N}) \\ &\leq \frac{1}{m} \lambda_{\max, S}(\mathbf{N}) \leq \eta \left(1 + c_R \sqrt{\frac{|S| \log(4/\delta)}{n}} + c_R \frac{|S| + \log(4/\delta)}{m} \right), \end{aligned}$$

where c_R is an universal constant. Assuming that $\mathbf{a}_3 |\theta_0^*| + \log(4/\delta)$ is small compared to m , we deduce that the spectrum of \mathbf{N} lies in $[1/2, 2]$. Thus, under \mathcal{B}_{1-1} , the spectrum of $m^{-1}(\mathbf{X}_S^{(0)T} \mathbf{X}_S^{(0)})$ restricted to its coordinates $S \times S$ lies in $[\eta^{-1}/2, 2\eta]$.

Turning to (54), we observe that $\mathbf{X}_S^{(0)} \theta_S^*$ follows a mean zero normal distribution. Using a deviation inequality for χ^2 distribution (Lemma 3), we deduce the existence of an event \mathcal{B}_{1-2} of probability larger higher $1 - \delta/4$ such that $\|\frac{1}{\sqrt{m}} \mathbf{X}_S^{(0)} \theta_S^*\|_2 \leq \|\theta_S^*\|_2 \sqrt{\eta} [1 + \sqrt{2 \log(4/\delta)/m}] \leq \|\theta_S^*\|_2 \sqrt{2\eta \log(4/\delta)}$, since m is large enough as assumed in Theorem 3. So from Equation (53), we deduce that, on $\mathcal{B}_1 = \mathcal{B}_{1-1} \cap \mathcal{B}_{1-2}$, we have

$$\left\| \frac{1}{m} \mathbf{X}_S^{(0)T} \mathbf{X}_S^{(0)} \theta_S^* \right\|_2 \leq \lambda_{\max}^{1/2} \left(\frac{1}{\sqrt{m}} \mathbf{X}_S^{(0)T} \mathbf{X}_S^{(0)} \right) \left\| \frac{1}{\sqrt{m}} \mathbf{X}_S^{(0)} \theta_S^* \right\|_2 \leq 2\eta \sqrt{\log(4/\delta)} \|\theta_S^*\|_2.$$

□

Proof of Lemma 10. $\hat{\sigma}_S^2 / \text{Var}(Y|\mathbf{X}_S)$ follows a χ^2 distribution with m degrees of freedom. Using a deviation inequality for χ^2 distribution (Lemma 3), we derive that $\hat{\sigma}_S^2 / \text{Var}(Y|\mathbf{X}_S) \in (1/2, 2)$, with probability higher than $1 - e^{-cm} \geq 1 - \delta/6$. Thus, it remains to bound $\text{Var}(Y|\mathbf{X}_S)$. From the definition of the property $\mathbf{S}[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$, we deduce that

$$\sigma^2 \leq \text{Var}(Y|\mathbf{X}_S) \leq \sigma^2 + \text{Var}(\mathbf{X}\theta_S^*) \leq \eta \|\theta_S^*\|_2^2 \leq \sigma^2 \left[1 + \eta \mathbf{a}_3^2 M(\mathbf{a}_1, \frac{\theta^*}{\sigma}) \frac{\log(p)}{m} \right].$$

□

5.5.2 Proof of Proposition 8

Write $\hat{\theta}_{MCP, N}$ for a stationary point of the MCP criterion and let \hat{S}_{MCP} denote its support. Since, we used the normalized design, we are more interested in the rescaled estimator $\hat{\theta}_{MCP}$ defined by $(\hat{\theta}_{MCP})_i = (\hat{\theta}_{MCP, N})_i / \|\mathbf{X}_{:,i}^{(1)}\|_2$. As explained in the proof of Lemma 10, the design matrix \mathbf{T} satisfies, with probability higher than $1 - 1/p$ the compatibility property (see [36, 42]) with any set of size less than $n/[c_\eta^{(1)} \log(p)]$, see [40]. Besides, the restricted eigenvalue condition satisfied for sparsities of size less than $n/[c_\eta^{(1)} \log(p)]$ are bounded by some constants depending on η , see [21, 49] with probability higher than $1 - 1/p$. From Lemma 1, we deduce that $\hat{\sigma}_{SL}/\sigma \in (3/4, 5/4)$ with probability higher than $1 - 1/p$. We are therefore in position to apply Theorem 6 in [49] and

Corollary 1 in Feng and Zhang [24]¹ provided that we chose the constant $c_\eta^{(MCP)}$ large enough and $c_\eta'^{(MCP)}$ small enough. From Theorem 6 in [49] with $B = S^*$ (the support of θ^*), we deduce that, with probability higher than $1 - 1/p$,

$$|\widehat{S}_{MCP}| \leq |\widehat{S}_{MCP} \setminus S^*| + |S^*| \leq c_\eta \|\theta^*\|_0 .$$

Write $\widehat{\theta}_{ls, S^*}^{(1)}$ for the least-square estimator of θ^* restricted to S^* : $\widehat{\theta}_{ls, S^*}^{(1)} = \arg \min_{\theta: S(\theta) \subset S^*} \|Y^{(1)} - \mathbf{X}^{(1)}\theta\|_2^2$ as defined in Equation (26). We deduce from Corollary 1 in [24] that

$$\|(\widehat{\theta}_{ls, S^*}^{(1)})_{\overline{S}_{MCP}}\|_2^2 \leq \|\widehat{\theta}_{MCP} - \widehat{\theta}_{ls, S^*}^{(1)}\|_2^2 \leq c_\eta'' \sigma^2 \frac{\log(p)}{m} M(c_\eta', \frac{\theta^*}{\sigma}) ,$$

The restricted least-square estimator $\widehat{\theta}_{ls, S^*}^{(1)}$ follows a normal distributions with mean θ^* and covariance $(\mathbf{X}_{S^*}^{(1)T} \mathbf{X}_{S^*}^{(1)})^{-1}$ where we consider here the pseudo-inverse. The eigenvalues of $m(\mathbf{X}_{S^*}^{(1)T} \mathbf{X}_{S^*}^{(1)})^{-1}$ are bounded by the restricted eigenvalue condition on the design $\mathbf{X}^{(1)}$. Hence, we obtain $\|\widehat{\theta}_{ls, S^*}^{(1)} - \theta^*\|_\infty \leq c_\eta''' \sigma \sqrt{\log(p)/m}$ with probability higher than $1 - c/p$, from some $c_\eta''' > 0$. This implies that $|\theta_i^*| \leq 2|(\widehat{\theta}_{ls, S^*}^{(1)})_i|$ if $|\theta_i^*| \geq 2c_\eta''' \sigma \sqrt{\log(p)/m}$. We obtain

$$\begin{aligned} \|\theta_{\overline{S}_{MCP}}^*\|_2^2 &\leq \sum_{i \in \overline{S}_{MCP}} \mathbf{1}_{|\theta_i^*| \geq 2c_\eta''' \sigma \sqrt{\log(p)/m}} 4|(\widehat{\theta}_{ls, S^*}^{(1)})_i|^2 + \mathbf{1}_{|\theta_i^*| \leq 2c_\eta''' \sigma \sqrt{\log(p)/m}} 2 \left(|(\widehat{\theta}_{ls, S^*}^{(1)})_i|^2 + (c_\eta''')^2 \sigma^2 \frac{\log(p)}{m} \right) \\ &\leq 4\|(\widehat{\theta}_{ls, S^*}^{(1)})_{\overline{S}_{MCP}}\|_2^2 + 2(c_\eta''')^2 \sigma^2 \frac{\log(p)}{m} M(c_\eta''', \frac{\theta^*}{\sigma}) \\ &\leq 4[c_\eta'' + (c_\eta''')^2] \sigma^2 \frac{\log(p)}{m} M(c_\eta' \vee c_\eta''', \frac{\theta^*}{\sigma}) . \end{aligned}$$

The result follows.

5.5.3 Proof of Theorem 4

To alleviate the notation, we simply write S_t for \widehat{S}_t in this proof. For a random vector $X \sim \mathcal{N}(0, \Sigma)$, we write $\Sigma^{(t)}$ for the conditional variance of X given $(X_{S_{t-1}}, S_{t-1})$. Standard computations for conditional variance based on Schur complement lead to

$$\Sigma^{(t)} = \Sigma_{\overline{S}_{t-1}, \overline{S}_{t-1}} - \Sigma_{\overline{S}_{t-1}, S_{t-1}} (\Sigma_{S_{t-1}, S_{t-1}})^{-1} \Sigma_{S_{t-1}, \overline{S}_{t-1}} , \quad (61)$$

where $(\Sigma_{S_{t-1}, S_{t-1}})^{-1}$ is the pseudo-inverse of $\Sigma_{S_{t-1}, S_{t-1}}$ obtained by considering its inverse when restricted to $S_{t-1} \times S_{t-1}$ and setting all its remaining entries to 0.

In the sequel, we denote $Y_\perp^{(t)} = \mathbf{\Pi}_{t, S_{t-1}}^\perp Y^{(t)}$ and $\mathbf{X}_\perp^{(t)} = \mathbf{\Pi}_{t, S_{t-1}}^\perp \mathbf{X}^{(t)}$. The following lemma ensures that the linear regression of $Y^{(t)}$ on $\mathbf{X}_\perp^{(t)}$ involves the restriction of θ^* to \overline{S}_{t-1} .

Lemma 11. *Fix any $t \in [1; T]$ and consider the event such that $|S_{t-1}| < m/T$. Then, given S_{t-1} , the rows of $\mathbf{X}_\perp^{(t)}$ are independent and follow a centered normal distribution with covariance matrix $\Sigma^{(t)}$. Besides, we have*

$$\left(Y_\perp^{(t)} | \mathbf{X}_\perp^{(t)}, S_{t-1} \right) \sim \mathcal{N}(\mathbf{X}_\perp^{(t)} \theta_{\overline{S}_{t-1}}^*, \sigma^2 \mathbf{I}_{m/T - |S_{t-1}|}) .$$

¹Actually, our definition of MCP uses a different normalization from that in [49] and [24] and one has therefore to translate their results in our setting.

The next lemma ensures that the population covariance matrix of the projected design still belongs to $\mathcal{U}[\eta]$.

Lemma 12. *For any $\Sigma \in \mathcal{U}(\eta)$ and any set S_{t-1} , The restriction of $\Sigma^{(t)}$ to $\bar{S}_{t-1} \times \bar{S}_{t-1}$ belongs to $\mathcal{U}(\eta)$.*

Denote $\delta = p^{-2}$. For $1 \leq t \leq T$, **Property (\mathbf{Q}_t)** is said to be satisfied if there exists an event ξ_t measurable with respect to $((\underline{Y}^{(1)}, \underline{X}^{(1)}), \dots, (\underline{Y}^{(t)}, \underline{X}^{(t)}))$ of probability higher than $(1 - \delta)^t$ such that the three following inequalities hold:

$$|\mathcal{S}(\theta^*) \setminus S_t| \leq \frac{\|\theta^*\|_0}{2^t} \vee 2M \left[2\sqrt{c_\eta T \frac{\log(p/\delta)}{\log(p)}}, \frac{\theta^*}{\sigma} \right]; \quad |S_t| \leq 2t\|\theta^*\|_0; \quad \|\theta_{\bar{S}_t}^*\|_2^2 \leq \sigma^2 c_\eta \|\theta_{\bar{S}_{t-1}}^*\|_0 \frac{T \log(p/\delta)}{m}. \quad (62)$$

Assume that the property \mathbf{Q}_T holds and recall that $\hat{S} = S_T$. Since $|\mathcal{S}(\theta^*) \setminus S_T|$ is an integer, $T \geq \log_2(n)$, and $\|\theta^*\|_0 \leq n/4$, there exists an event of probability larger than $(1 - \delta)^T$ such that

$$\begin{aligned} |\mathcal{S}(\theta^*) \setminus \hat{S}| &\leq 2M \left[2\sqrt{c_\eta T \frac{\log(p/\delta)}{\log(p)}}, \frac{\theta^*}{\sigma} \right]; \quad |\hat{S}| \leq 2T\|\theta^*\|_0; \\ \|\theta_{\hat{S}}^*\|_2^2 &\leq \sigma^2 c_\eta 2M \left[2\sqrt{c_\eta T \frac{\log(p/\delta)}{\log(p)}}, \frac{\theta^*}{\sigma} \right] \frac{T \log(p/\delta)}{m}, \end{aligned}$$

which, with $\delta = p^{-2}$, is the result of Theorem 4. Thus, it suffices to prove (\mathbf{Q}_t) by induction.

Lemma 13. *Assume that $\Sigma \in \mathcal{U}(\eta)$ with $\eta > 0$. Assume that $|S_t|$ is such that*

$$\underline{c}_\eta^{(SL),2} [\|\theta^*\|_0 \log(p/\delta) + \log(1/\delta) \log(p)] \leq m/T - |S_t|.$$

(Recall that $\underline{c}_\eta^{(SL),2}$ is introduced in Lemma 1). Then, given S_t , there exists an event \mathcal{F}_{t+1} measurable with respect to $(\underline{Y}^{(t+1)}, \underline{X}^{(t+1)})$ of probability higher than $1 - \delta$ such that

$$\|\theta_{\bar{S}_{t+1}}^*\|_2^2 \leq c_\eta \sigma^2 \frac{T \|\theta_{\bar{S}_t}^*\|_0}{m} \log\left(\frac{p}{\delta}\right) \quad \text{and} \quad |S_{t+1}| \leq |S_t| + 2\|\theta_{\bar{S}_t}^*\|_0.$$

Step (\mathbf{Q}_1) : Recall that $S_0 = \emptyset$. By Lemma 13 and Equation (30), there exists an event \mathcal{E}_1 with probability higher than $1 - \delta$ such that

$$\|\theta_{\bar{S}_1}^*\|_2^2 \leq c_\eta \sigma^2 \|\theta^*\|_0 \frac{T \log(p/\delta)}{m} \quad \text{and} \quad |S_1| \leq 2\|\theta^*\|_0.$$

Counting the components of $\theta_{\bar{S}_1}^*$ that are larger (in absolute value) than $2\sigma\sqrt{c_\eta T \log(p/\delta)/m}$, we derive that

$$\|\theta_{\bar{S}_1}^*\|_2^2 \geq \left[|\mathcal{S}(\theta^*) \setminus S_1| - M \left[2\sqrt{c_\eta T \frac{\log(p/\delta)}{\log(p)}}, \frac{\theta^*}{\sigma} \right] \right] 4c_\eta \sigma^2 \frac{T}{m} \log\left(\frac{p}{\delta}\right),$$

which, together with the previous bound implies

$$|\mathcal{S}(\theta^*) \setminus S_1| \leq M \left[2\sqrt{c_\eta T \frac{\log(p/\delta)}{\log(p)}}, \frac{\theta^*}{\sigma} \right] + \frac{\|\theta^*\|_0}{4} \leq \frac{\|\theta^*\|_0}{2} \vee \left(2M \left[2\sqrt{c_\eta T \frac{\log(p/\delta)}{\log(p)}}, \frac{\theta^*}{\sigma} \right] \right).$$

So (\mathbf{Q}_1) holds.

Induction step: Assume that (\mathbf{Q}_{t-1}) holds for some $T-1 \geq t \geq 1$. By (\mathbf{Q}_{t-1}) and on ξ_{t-1} , we have that $|S_{t-1}| \leq 2(t-1)\|\theta^*\|_0 \leq m/(2T)$ by Condition (30). Thus, $m/T - |S_{t-1}|$ is large enough and we can apply Lemma 13. As a consequence, there exists an event \mathcal{E}_t of probability higher than $(1-\eta)^t$ such that

$$\|\theta_{\bar{S}_t}^*\|_2^2 \leq c_\eta \sigma^2 \frac{T\|\theta_{\bar{S}_{t-1}}^*\|_0}{m} \log\left(\frac{p}{\delta}\right) \quad \text{and} \quad |S_t| \leq |S_{t-1}| + 2\|\theta_{\bar{S}_{t-1}}^*\|_0.$$

Together with (\mathbf{Q}_{t-1}) , this implies $|S_t| \leq 2(t-1)\|\theta^*\|_0 + 2\|\theta^*\|_0 = 2t\|\theta^*\|_0$. As for the proof of (\mathbf{Q}_1) , we lower bound $\|\theta_{\bar{S}_t}^*\|_2^2$ by considering separately the entries larger than (in absolute value) than $2\sigma\sqrt{c_\eta T \log(p/\delta)/m}$. This leads us to

$$\begin{aligned} |\mathcal{S}(\theta^*) \setminus S_t| &\leq M \left[2\sqrt{c_\eta T \frac{\log(p/\delta)}{\log(p)}}, \frac{\theta_{\bar{S}_{t-1}}^*}{\sigma} \right] + \frac{\|\theta_{\bar{S}_{t-1}}^*\|_0}{4} \leq \frac{\|\theta_{\bar{S}_{t-1}}^*\|_0}{2} \vee \left(2M \left[2\sqrt{c_\eta T \frac{\log(p/\delta)}{\log(p)}}, \frac{\theta^*}{\sigma} \right] \right) \\ &\leq \frac{\|\theta^*\|_0}{2^t} \vee \left(2M \left[2\sqrt{c_\eta T \frac{\log(p/\delta)}{\log(p)}}, \frac{\theta^*}{\sigma} \right] \right), \end{aligned}$$

where we used (\mathbf{Q}_{t-1}) in the second line. We have proved (\mathbf{Q}_t) . This concludes the proof.

Proof of Lemma 11. To alleviate the notation, we simply write S for S_{t-1} , \hat{S} for $\hat{S}^{(ith)}$, \mathbf{X} (resp. Y) for $\underline{\mathbf{X}}^{(t)}$ (resp. $\underline{Y}^{(t)}$), \mathbb{E} for the expectation $\mathbb{E}^{(t)}$, and $\underline{\Pi}_S^\perp$ for $\underline{\Pi}_{t,S}^\perp$ (in the proof of this lemma only). Besides, since S has been built based on independent samples, we consider it as fixed. Also, without loss of the generality, we assume that $S = \{1, \dots, |S|\}$.

Define $\mathbf{Z} = \mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{X}_S]$. Since \mathbf{X} follows a normal distribution, \mathbf{Z} is independent of \mathbf{X}_S . Besides, the rows of \mathbf{Z} are i.i.d. distributed according to centered normal distribution with covariance matrix $\Sigma^{(t)}$. Since the rows of \mathbf{X} are i.i.d., each column of $\mathbb{E}[\mathbf{X}|\mathbf{X}_S]$ is a linear combination of the columns of \mathbf{X}_S . As a consequence, there exists a $|S| \times p$ matrix \mathbf{R} such that $\mathbb{E}[\mathbf{X}|\mathbf{X}_S] = \mathbf{X}_S \mathbf{R}$.

Since $T|S| < m$ and since Σ is invertible, the rank of $V[S, \mathbf{X}]$ equals $|S|$ almost surely. As a consequence, applying the orthogonal projection along $V[S, \mathbf{X}]$ to \mathbf{X} leads to

$$\underline{\Pi}_S^\perp \mathbf{X} = \underline{\Pi}_S^\perp \mathbf{Z} + \underline{\Pi}_S^\perp \mathbf{X}_S \mathbf{R} = \underline{\Pi}_S^\perp \mathbf{Z}.$$

Since the rows of \mathbf{Z} are i.i.d. with covariance $\Sigma^{(t)}$, there exists a matrix \mathbf{U} with i.i.d. standard normal entries such that $\mathbf{Z} = \mathbf{U}\mathbf{\Gamma}^{(t)}$ where $\mathbf{\Gamma}$ is a square root of $\Sigma^{(t)}$. As a consequence, $\underline{\Pi}_S^\perp \mathbf{X} = \underline{\Pi}_S^\perp \mathbf{U}\mathbf{\Gamma}$. Since \mathbf{X}_S is independent of \mathbf{U} it follows that, given \mathbf{X}_S , the $(m/T - |S|) \times p$ matrix $\underline{\Pi}_S^\perp \mathbf{U}$ is made of independent standard normal entries and the rows of $\mathbf{X}_\perp^{(t)} = \underline{\Pi}_S^\perp \mathbf{X}$ therefore follow independent normal distributions with covariance matrix $\Sigma^{(t)}$.

Also we have

$$Y_\perp^{(t)} = \underline{\Pi}_S^\perp Y = \mathbf{X}_\perp^{(t)} \theta^* + \sigma \underline{\Pi}_S^\perp \epsilon = \mathbf{X}_\perp^{(t)} \theta_S^* + \sigma \underline{\Pi}_S^\perp \epsilon,$$

since the columns of $\mathbf{X}_\perp^{(t)}$ in S are equal to zero. Given \mathbf{X}_S , $\underline{\Pi}_S^\perp \epsilon$ is projection of a standard normal vector onto a subspace of dimension $m/T - |S|$. As a consequence, $\underline{\Pi}_S^\perp \epsilon$ follows a normal distribution with covariance matrix $\mathbf{I}_{m/T - |S|}$ and is independent of \mathbf{X} . The result follows. \square

Proof of Lemma 12. For simplicity, we write S for S_{t-1} . Let u be a normed vector supported in \bar{S} . We shall prove that $u^T \Sigma^{(t)} u$ belongs to $(1/\eta, \eta)$. Consider a random vector $X \sim \mathcal{N}(0, \Sigma)$ so that $u^T \Sigma^{(t)} u = \text{Var}(u^T X | X_S)$. Consider the $|S| + 1$ size covariance matrix $\mathbf{\Gamma}$ of $((X_i)_{i \in S}, u^T X)$. Then, $\mathbf{\Gamma} \in \mathcal{U}(\eta)$ and $\text{Var}(u^T X | \mathbf{X}_S) = 1/(\mathbf{\Gamma}_{|S|+1, |S|+1}^{-1})$, which therefore lies in $(1/\eta, \eta)$. \square

Proof of Lemma 13. To alleviate the notation, we simply write $\hat{\theta}_{(SL)}$ and $\hat{\theta}_{(SL,t)}$ for $\hat{\theta}_{(SL)}[\underline{Y}_{\perp}^{(t+1)}, \underline{\mathbf{X}}_{\perp}^{(t+1)}]$ and $\hat{\theta}_{(SL,t)}[\underline{Y}_{\perp}^{(t+1)}, \underline{\mathbf{X}}_{\perp}^{(t+1)}]$ respectively. Recall that $S_{t+1} = S_t \cup \mathcal{S}(\hat{\theta}_{(SL,t)})$. The rows of $\underline{\mathbf{X}}_{\perp}^{(t+1)}$ corresponding to indices in S_t are null. Therefore, $\hat{\theta}_{(SL)}[\underline{Y}_{\perp}^{(t+1)}, \underline{\mathbf{X}}_{\perp}^{(t+1)}]$ is a square-root Lasso estimator of $\underline{Y}_{\perp}^{(t+1)}$ given the restriction of $\underline{\mathbf{X}}_{\perp}^{(t+1)}$ to the rows in \overline{S}_t . In view of Lemmas 11 and 12, we can apply Lemma 1. Thus, given S_t , there exists an event \mathcal{F}_t of probability higher than $1 - \delta$ such that

$$\|\theta_{\overline{S}_t}^* - \hat{\theta}_{(SL)}\|_2^2 \leq \underline{c}_{\eta}^{(SL)} \sigma^2 \frac{\|\theta_{\overline{S}_t}^*\|_0}{m/T - |S_{t-1}|} \log\left(\frac{p}{\delta}\right) \quad \text{and} \quad \left|\frac{\widehat{\sigma}^{(t)}}{\sigma} - 1\right| \leq \frac{1}{4}, \quad (63)$$

By assumption, $m/T \geq 2|S_{t-1}|$. Since $\hat{\theta}_{(SL,t)}$ is a hard thresholded modification of $\hat{\theta}_{(SL)}$ at level

$$\frac{8}{3} \widehat{\sigma}^{(t)} \sqrt{\underline{c}_{\eta}^{(SL)} \frac{T}{m} \log\left(\frac{p}{\delta}\right)} \geq 2\sigma \sqrt{\underline{c}_{\eta}^{(SL)} \frac{T}{m} \log\left(\frac{p}{\delta}\right)},$$

its entry-wise error increases only at the non-zero entries of $\theta_{\overline{S}_t}^*$ and at most by $10/3\sigma \sqrt{\underline{c}_{\eta}^{(SL)} \frac{T}{m} \log\left(\frac{p}{\delta}\right)}$. This implies that

$$\|\theta_{\overline{S}_{t+1}}^*\|_2^2 \leq \|\theta_{\overline{S}_t}^* - \hat{\theta}_{(SL,t)}\|_2^2 \leq 2\left(2 + \frac{100}{9}\right) \sigma^2 \underline{c}_{\eta}^{(SL)} \frac{T \|\theta_{\overline{S}_t}^*\|_0}{m} \log\left(\frac{p}{\delta}\right).$$

Recall that $(S_{t+1} \setminus S_t)$ is the support of $\hat{\theta}_{(SL,t)}$. Each non-zero entry of $\hat{\theta}_{(SL,t)}$ is equal to that of $\hat{\theta}_{(SL)}$. As a consequence, each index in the support of $\hat{\theta}_{(SL,t)}$ and outside the support of θ^* contributes at least by $2\sigma^2 \underline{c}_{\eta}^{(SL)} \frac{T}{m} \log\left(\frac{p}{\delta}\right)$ in the loss $\|\theta_{\overline{S}_t}^* - \hat{\theta}_{(SL)}\|_2^2$. This implies

$$\|\theta_{\overline{S}_t}^* - \hat{\theta}_{(SL)}\|_2^2 \geq 2\sigma^2 \underline{c}_{\eta}^{(SL)} \frac{T}{m} \log\left(\frac{p}{\delta}\right) |S_{t+1} \setminus \mathcal{S}(\theta_{\overline{S}_t}^*)|,$$

which in view of (63) leads us to $|S_{t+1} \setminus \mathcal{S}(\theta_{\overline{S}_t}^*)| \leq \|\theta_{\overline{S}_t}^*\|_0$ and

$$|S_{t+1} \setminus S_t| \leq \|\theta_{\overline{S}_t}^*\|_0 + |S_{t+1} \setminus \mathcal{S}(\theta_{\overline{S}_t}^*)| \leq 2\|\theta_{\overline{S}_t}^*\|_0,$$

which concludes the proof. □

6 Proofs of the minimax lower bounds

We first state the following classical lemma that links the total variation distance with the performance of a test with composite hypotheses. Some variants of it may be found in textbooks such as [43]. For a sake of completeness, we provide a proof below.

Lemma 14. *Consider a parametric model $\{\mathbb{P}_{\theta}, \theta \in \Theta\}$ and two subsets $\Theta_0 \subset \Theta, \Theta_1 \subset \Theta$. Let μ_0 and μ_1 be any probability measures on Θ . Denote $\mathbf{P}_{\mu_i} = \int \mathbb{P}_{\theta} \mu_i(d\theta)$ for $i = 0, 1$. Any test ϕ of Θ_0 against Θ_1 satisfies*

$$\sup_{\theta \in \Theta_0} \mathbb{P}_{\theta}[\phi = 1] + \sup_{\theta \in \Theta_1} \mathbb{P}_{\theta}[\phi = 0] \geq 1 - \mu_0[\theta \notin \Theta_0] - \mu_1[\theta \notin \Theta_1] - \|\mathbf{P}_{\mu_0} - \mathbf{P}_{\mu_1}\|_{TV}. \quad (64)$$

Proof of Lemma 14. For $i = 0, 1$, define the probability measure μ'_i by $\mu'_i[A] = \mu_i[A \cap \Theta_i]/\mu_i[\Theta_i]$ for any event A . Given μ'_i , let $\mathbf{P}'_{\mu_i} = \int \mathbb{P}_\theta \mu'_i(d\theta)$. It follows from Le Cam's arguments that

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\phi = 1] + \sup_{\theta \in \Theta_1} \mathbb{P}_\theta[\phi = 0] \geq 1 - \|\mathbf{P}'_{\mu_0} - \mathbf{P}'_{\mu_1}\|_{TV}. \quad (65)$$

By triangular inequality, one has

$$\|\mathbf{P}'_{\mu_0} - \mathbf{P}'_{\mu_1}\|_{TV} \leq \|\mathbf{P}_{\mu_0} - \mathbf{P}_{\mu_1}\|_{TV} + \|\mathbf{P}'_{\mu_0} - \mathbf{P}_{\mu_0}\|_{TV} + \|\mathbf{P}'_{\mu_1} - \mathbf{P}_{\mu_1}\|_{TV}$$

Obviously, the total variation distance $\|\mu'_0 - \mu_0\|_{TV}$ equals $\mu'_0[\Theta_0] - \mu_0[\Theta_0] = \mu_0[\overline{\Theta_0}]$.

$$\|\mathbf{P}'_{\mu_0} - \mathbf{P}_{\mu_0}\|_{TV} = \sup_{\mathcal{A}} \left| \int \mathbb{P}_\theta(\mathcal{A})[\mu'_0(d\theta) - \mu_0(d\theta)] \right| \leq \|\mu'_0 - \mu_0\|_{TV}.$$

Arguing similarly for $\|\mathbf{P}'_{\mu_1} - \mathbf{P}_{\mu_1}\|_{TV}$ and plugging these bound into (65) concludes the proof. \square

6.1 Proof of Proposition 1

Proof of Proposition 1. Intuitively, testing the sparsity for $k_0 \geq n$ is impossible because θ^* cannot be even recovered in noiseless setting ($\sigma = 0$) when it contains more than n non-zero entries. As the design matrix \mathbf{X} is random, this argument needs to be slightly refined. Without loss of generality, we consider the case $p = n + 1$, $k_0 = n$ and $\Delta = 1$. Let us write $\underline{\mathbf{X}}$ the submatrix of \mathbf{X} made of its n first columns. In order to apply Lemma 14, we shall build two suitable prior distributions on the set of n and $n + 1$ sparse vectors.

With probability one, the square matrix $\underline{\mathbf{X}}$ is invertible. Also denote s_{\min} (resp. s_{\max}) the smallest (resp. highest) singular values of $\underline{\mathbf{X}}$. Fix any $\delta \in (0, 1)$. As stated for instance in [41], there exist $c_-(n, \delta) = c_- > 0$, $c_+(n, \delta) = c_+ > 0$ such that the following holds

$$\mathbb{P}_{\mathbf{X}}[s_{\min} > c_- ; s_{\max} < c_+ ; c_- < \|\mathbf{X}_{:,p}\|_2 < c_+] \geq 1 - \delta, \quad (66)$$

where $\mathbb{P}_{\mathbf{X}}$ stands for the distribution of \mathbf{X} . Here, $\mathbf{X}_{:,p}$ stands for the p -th column of \mathbf{X} . Although the exact expression of c_- and c_+ is not relevant in this proof, these two quantities are of the order $n^{-1/2}$ and $n^{1/2}$. We call \mathcal{A} the event defined in the above probability bound.

Let μ_0 stand for the centered Gaussian measure in \mathbb{R}^{n+1} with covariance matrix $\begin{pmatrix} \mathbf{I}_n & 0 \\ 0 & 0 \end{pmatrix}$.

We write $\mathbf{P}_{0,0} = \int_{\mathbb{R}^{n+1}} \mathbb{P}_{\theta,0} \mu_0(d\theta)$. Given any $r > 0$, define the vector $v_r = (0, \dots, 0, r)^T$. We fix $\mathbf{P}_{1,r,0} = \int_{\mathbb{R}^{n+1}} \mathbb{P}_{\theta+v_r,0} \mu_0(d\theta)$. We argue that, for r small enough, the total variation distance $\|\mathbf{P}_{0,0} - \mathbf{P}_{1,r,0}\|_{TV}$ is smaller than 2δ .

Under $\mathbf{P}_{0,0}$, for a fixed \mathbf{X} , it holds that $Y \sim \mathcal{N}(0, \underline{\mathbf{X}}\underline{\mathbf{X}}^T)$ whereas, under $\mathbf{P}_{1,r,0}$, it holds that $Y \sim \mathcal{N}(\mathbf{X}v_r, \underline{\mathbf{X}}\underline{\mathbf{X}}^T)$. When $\underline{\mathbf{X}}$ satisfies \mathcal{A} , these two covariance matrices are invertible with eigenvalues in (c_-^2, c_+^2) and $\|\mathbf{X}v_r\|_2 \leq rc_+$. Thus, for r going to zero, the total variation distance between these conditional distributions goes to zero uniformly over all \mathbf{X} satisfying \mathcal{A} . In particular, there exists some r_0 such that these distances are uniformly smaller than δ . Since $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$, it follows that

$$\|\mathbf{P}_{0,0} - \mathbf{P}_{1,r_0,0}\|_{TV} \leq 2\delta.$$

Consider $\sigma_0 > 0$ whose value will be fixed later. Define $\mathbf{P}_{0,\sigma_0} = \int_{\mathbb{R}^{n+1}} \mathbb{P}_{\theta,\sigma_0} \mu_0(d\theta)$ and $\mathbf{P}_{1,r_0,\sigma_0} = \int_{\mathbb{R}^{n+1}} \mathbb{P}_{\theta+v_r,\sigma_0} \mu_0(d\theta)$ the distributions associated to the linear regression models. By contraction properties of the total variation distances, one has

$$\|\mathbf{P}_{0,\sigma_0} - \mathbf{P}_{1,r_0,\sigma_0}\|_{TV} \leq \|\mathbf{P}_{0,0} - \mathbf{P}_{1,r_0,0}\|_{TV} \leq 2\delta.$$

When θ is sampled according to μ_0 , then the smallest (in absolute value) entry of θ among the n first entries is larger than some positive quantity \underline{c}_- , with probability larger than $1 - \delta$. Let us call \mathcal{B} the corresponding event. Define $\underline{\mu}$ as the measure μ_0 conditioned to the event \mathcal{B} , i.e. $\underline{\mu}(\mathcal{C}) = \mu_0(\mathcal{C} \cap \mathcal{B})/\mu_0(\mathcal{B})$ for any measurable event \mathcal{C} . Then, we introduce $\underline{\mathbf{P}}_{1,r_0,\sigma_0} = \int_{\mathbb{R}^{n+1}} \mathbb{P}_{\theta+v_r,\sigma_0} \underline{\mu}(d\theta)$. By triangular inequality, we obtain

$$\|\mathbf{P}_{0,\sigma_0} - \underline{\mathbf{P}}_{1,r_0,\sigma_0}\|_{TV} \leq 3\delta .$$

When θ is sampled according to $\underline{\mu}$, $(\theta+v_{r_0})$ satisfies $d_2(\theta+v_{r_0}, \mathbb{B}_0[n]) \geq \underline{c}_- \wedge r_0$. As a consequence of Lemma 14, any test of $\{\|\theta\|_0 \leq n, \sigma = \sigma_0\}$ versus $\{\|\theta\|_0 \leq n+1, d_2(\theta+v_{r_0}, \mathbb{B}_0[n]) \geq \underline{c}_- \wedge r_0, \sigma = \sigma_0\}$ has a risk higher than $1 - 3\delta$. We have

$$\rho_{3\delta}^*[k_0, \Delta] \geq \frac{\underline{c}_- \wedge r_0}{\sigma_0} ,$$

where $\underline{c}_- \wedge r_0$ does not depend on σ_0 . Taking σ_0 arbitrarily small leads to the desired result. \square

6.2 Proof of Theorem 1

Given integers k_0 and $\Delta \leq p - k_0$, and $\rho > 0$, we define the collection

$$\mathbb{B}_0[k_0, \Delta, \rho] = \mathbb{B}_0[k_0 + \Delta] \cap \{\theta : d_2(\theta; \mathbb{B}_0[k_0]) \geq \rho\} .$$

We start by a simple reduction result to narrow the range of parameters. Its proof is postponed to the end of the section.

Lemma 15. *For any $\Delta' \leq \Delta \leq p - k_0$, we have*

$$\rho_\gamma^*[k_0, \Delta] \geq \rho_\gamma^*[k_0, \Delta'] . \quad (67)$$

For the sake of the following bound, we explicit the dependency of $\rho_\gamma^[k_0, \Delta]$ with respect to p by denoting it $\rho_\gamma^*[p, k_0, \Delta]$. For any $k'_0 < k_0 < p$ and $\Delta \leq p - k_0$, we have*

$$\rho_\gamma^*[p, k_0, \Delta] \geq \rho_\gamma^*[p - k_0 + k'_0, k'_0, \Delta] . \quad (68)$$

In other words, the minimax separation distance is non-decreasing with respect to Δ and, up to a change in the number p of covariates, it is also nondecreasing with respect to k_0 . Next, we state three lemmas whose combination implies Theorem 1.

Lemma 16. *Assume that $p \geq 2n$. There exists a numerical constant $c > 0$ such that*

$$\rho_\gamma^{*2}[k_0, \Delta] \geq c \left[\frac{1}{\sqrt{n}} \wedge \frac{\Delta \log(1 + \sqrt{p}/\Delta)}{n} \right] , \quad (69)$$

for any $\gamma \leq 0.53$, all $k_0 \leq n$ and $1 \leq \Delta \leq p - k_0$.

Proof of Lemma 16. This lemma is a consequence of known signal detection lower bounds ($k_0 = 0$). For instance, it is proved in [46, Sect.9.1] in

$$\rho_\gamma^{*2}[0, \Delta] \geq c \left[\frac{1}{\sqrt{n}} \wedge \frac{\Delta \log(1 + \sqrt{p}/\Delta)}{n} \right] ,$$

for all $1 \leq \Delta \leq p$. Since $p \geq 2n$ and $k_0 \leq n$, Lemma 15 entails that

$$\rho_\gamma^{*2}[k_0, \Delta] \geq c \left[\frac{1}{\sqrt{n}} \wedge \frac{\Delta \log(1 + \sqrt{p/2}/\Delta)}{n} \right] ,$$

which concludes the proof. \square

Lemma 17. Assume that $p \geq 2n$. There exist constants c_1 – c_5 such that the following holds for all $\gamma \leq 0.06$:

$$\rho_\gamma^{*2}[k_0, \Delta] \geq c_1 \frac{\Delta \wedge k_0}{n} \log \left(1 + \frac{\sqrt{p}}{\Delta \wedge k_0} \right), \quad (70)$$

for all $k_0 \leq n$ and $\Delta > 0$. Furthermore, if $p \geq c_2 n^2$ and $k_0 \in (c_3 n / \log(\sqrt{p}/n), \sqrt{p}/e^4)$ and $\Delta \leq k_0$, then

$$\rho_\gamma^{*2}[k_0, \Delta] \geq c_4 \frac{\Delta}{n} \log \left(2 \vee \frac{\sqrt{p}}{k_0} \right) e^{c_5 \frac{k_0}{n} \log(1 + \frac{\sqrt{p}}{k_0})}. \quad (71)$$

Proof of Lemma 17. In the above lemma, the minimax lower bounds both depend on the size k_0 of the null hypothesis and on the size Δ of the alternative hypothesis. As a consequence, we cannot directly rely anymore on signal detection results as in the previous lemma. Nevertheless, we will introduce a third party hypothesis and make use of previous signal detection lower bounds for unknown σ [45, 46].

By Lemma 15, we assume without loss of generality that $\Delta \leq k_0$. Given $\rho > 0$ and $1 \leq k \leq p$, we define $\mu_{\rho,k}$ as the uniform measure over the set

$$\left\{ \theta \in \mathbb{R}^p, \|\theta\|_0 = k, \quad \forall i \in \{1, \dots, p\}, \quad |\theta_i| = 0 \text{ or } \rho/\sqrt{k} \right\},$$

and the mixture measure $\mathbf{P}_{\rho,k} = \int \mathbb{P}_{\theta,1} \mu_{\rho,k}(d\theta)$. As a way to derive minimax lower bounds for signal detection with unknown noise level, it is proved in [45, Theorem 4.3] and [46, Lemma 9.3]² that $\|\mathbf{P}_{\rho,k} - \mathbb{P}_{0,1+\rho^2}\|_{TV} \leq 0.47$ if $\rho \leq \rho_k$ or if $\rho \leq \rho'_k$ with

$$\begin{aligned} \rho_k^2 &= \frac{k}{2n} \log \left(1 + \frac{\sqrt{p}}{k} \right); \\ \rho'_k &= -1 + \left(\frac{p}{2ek} \right)^{k/n} (4k)^{-2/n} \quad \text{for} \quad k \log \left(\frac{\sqrt{p}}{e^{3/2}k} \right) \geq n. \end{aligned}$$

Let us now deduce (70). Since ρ_k is increasing with respect to k , we have

$$\|\mathbf{P}_{\rho_{k_0}, k_0} - \mathbf{P}_{\rho_{k_0}, k_0 + \Delta}\|_{TV} \leq \|\mathbf{P}_{\rho_{k_0}, k_0} - \mathbb{P}_{0,1+\rho_{k_0}^2}\|_{TV} + \|\mathbb{P}_{0,1+\rho_{k_0}^2} - \mathbf{P}_{\rho_{k_0}, k_0 + \Delta}\|_{TV} \leq 0.94$$

Under $\mu_{\rho_{k_0}, k_0}$, θ is k_0 -sparse, whereas under $\mu_{\rho_{k_0}, k_0 + \Delta}$, θ is $k_0 + \Delta$ -sparse and its square distance to $\mathbb{B}_0[k_0]$ is $\Delta \rho_{k_0}^2 / (k_0 + \Delta)$. From Lemma 14, we deduce that, for $\gamma \leq 0.06$, one has

$$\rho_\gamma^{*2}[k_0, \Delta] \geq \Delta \rho_{k_0}^2 \frac{\Delta}{\Delta + k_0} \geq \rho_{k_0}^2 \frac{\Delta}{2k_0},$$

which enforces (70) since we have $\Delta \leq k_0$. Turning to (71), we observe that, under the assumptions of the lemma (and with a suitable choice of c_2), $k \log \left(\frac{\sqrt{p}}{e^{3/2}k} \right) \geq n$ both for $k = k_0$ and $k = k_0 + \Delta \leq 2k_0$. Arguing as above, we deduce that

$$\rho_\gamma^{*2}[k_0, \Delta] \geq \rho_{k_0}^{\prime 2} \frac{\Delta}{2k_0} \geq c \frac{\Delta}{k_0} e^{c' \frac{k_0}{n} \log \left(\frac{p}{k_0} \right)} \geq c_2 \frac{\Delta}{n} \log \left(\frac{p}{k_0} \right) e^{c_3 \frac{k_0}{n} \log \left(\frac{p}{k_0} \right)},$$

since the expression inside the exponential is bounded away from zero and since $e^x \geq 1 + x$ for $x > 0$. We have proved (71). \square

²Actually, the results in [45, 46] are expressed in terms of minimax separation distance, the total variation distance control being stated in their respective proof.

The following lemma provides the key new lower bound. It corresponds to the regime where both k_0 and Δ are large. Its proof relies on more advanced arguments than the other regimes.

Lemma 18. *There exists positive numerical constant c and c_2 such that the following holds for any $\gamma \leq 0.5$ and all $p \geq c_2$. For any $p^{1/4} \leq k_0 \leq n$ and $\Delta \geq k_0^{2/3} \vee p^{1/4}$, one has*

$$\rho_\gamma^{*2}[k_0, \Delta] \geq c \frac{\Delta \log^2 [1 + \sqrt{\frac{k_0}{\Delta}}]}{n \log(p)}. \quad (72)$$

Proof of Theorem 1. First we prove (6). The case $k_0 \leq \sqrt{p}$ is a consequence of Lemmas 16 and 17. As for the case $k_0 \in (\sqrt{p}, n)$, we divide the analysis into several subcases. If $\Delta \leq p^{1/4}$, it follows from Lemma 17 that $\rho_\gamma^{*2}[k_0, \Delta]$ is at least of the order of $\Delta \log(p)/n$ which is larger than the lower bound in (6). For $\Delta \geq p^{1/4} \vee k_0^{2/3}$ we rely on Lemma 18. For $\Delta \in (p^{1/4}, k_0^{2/3})$, we define $k'_0 = \lfloor \Delta^{3/2} \rfloor$. From the reduction (68) and Lemma 18, we derive that

$$\rho_\gamma^{*2}[k_0, \Delta] \geq c \frac{\Delta \log^2 [1 + \sqrt{\frac{k'_0}{\Delta}}]}{n \log(p - k_0 + k'_0)} \geq c' \frac{\Delta}{n} \log(p) \geq c'' \frac{\Delta \log^2 [1 + \sqrt{\frac{k_0}{\Delta}}]}{n \log(p)}.$$

Finally, the lower bound (7) is a consequence of the second part of Lemma 17 together with the reduction lemma 15. □

Proof of Lemma 15. The first bound is a simple consequence of the inclusion $\mathbb{B}_0[k_0, \Delta, \rho] \subset \mathbb{B}_0[k_0, \Delta', \rho]$. Let us turn to (68). Take any $\zeta > 0$ arbitrarily small and define $r = \rho_\gamma^*[k_0, \Delta] + \zeta$. There exists a test ϕ satisfying $R[\phi; k_0, \Delta, r] \leq \gamma$. For any linear regression problem with $p - k_0 + k'_0$ covariates and response Y , we sample $k_0 - k'_0$ new independent covariates, write $\underline{\mathbf{X}}$ the corresponding new design matrix of size $n \times (k_0 - k'_0)$, and define $\underline{\mathbf{Y}} = Y + r \underline{\mathbf{X}} \mathbf{1}$ where $\mathbf{1}$ is the constant vector of size $k_0 - k'_0$. Since $R[\phi; k_0, \Delta, r] \leq \gamma$, we have

$$\sup_{\theta, \|\theta\|_0 \leq k'_0} \mathbb{P}_\theta[\phi(\underline{\mathbf{Y}}) = 1] + \sup_{\theta, \|\theta\|_0 \leq k'_0 + \Delta, d_2(\theta, \mathbb{B}_0[k'_0]) \geq r} \mathbb{P}_\theta[\phi(\underline{\mathbf{Y}}) = 0] \leq \gamma,$$

implying that $\rho_\gamma^*[p - k_0 + k'_0, k'_0, \Delta] \leq r$. Taking the infimum over all $\zeta > 0$, we obtain (68). □

Proof of Lemma 18. Without loss of generality we assume that the noise level σ is equal to one and we write \mathbb{P}_θ for $\mathbb{P}_{\theta, 1}$. Since the minimax separation distance $\rho_\gamma^*[k_0, \Delta]$ is a nondecreasing function of Δ , we have $\rho_\gamma^*[k_0, \Delta] \geq \rho_\gamma^*[k_0, k_0]$ for any $\Delta > k_0$. In view of (72) and since $\log(1 + x) \geq x/2$ for any $x \in [0, 1]$, we only need to prove (72) for $\Delta \leq k_0$.

Define $\bar{k}_0 = k_0 - \Delta/2$ and $\bar{k}_1 = k_0 + \Delta/2$. We introduce two priors $\mu_0^{\otimes p}$ and $\mu_1^{\otimes p}$ that are almost supported on $\mathbb{B}_0[k_0]$ and $\mathbb{B}_0[k_0 + \Delta]$ respectively and such that the first moments of μ_0 and μ_1 are matching. In Step 3 below, we show that this moment matching property ensures that the corresponding mixture distributions of (Y, \mathbf{X}) are close in total variation distance.

Step 1. Construction of the priors. As in [17], we build prior measures μ_0 and μ_1 in such a way that their first moments are matching. Define the two quantities where m is redefined only in this proof as follows)

$$m = 2 \lfloor 2 \log(p) \rfloor, \quad M = c \sqrt{\log(p)/n}, \quad (73)$$

for some universal constant c whose value will be fixed later. The following result is borrowed from [17, Lemma 3].

Lemma 19. *Given any positive and even integer m and $q \in (0, 1)$, define*

$$a_m = \tanh \left[\frac{1}{m} \arg \cosh \left(\frac{1+q}{1-q} \right) \right]. \quad (74)$$

There exists two positive and symmetric measures ν_0 and ν_1 whose support lie in $[-1, -a_m] \cup [a_m, 1]$ satisfying:

$$\int \nu_0(dt) = q; \quad \int \nu_1(dt) = 1; \quad (75)$$

$$\int t^d \nu_0(dt) = \int t^d \nu_1(dt); \quad d = 1, \dots, m. \quad (76)$$

Fix $q = \bar{k}_0/\bar{k}_1$. Then, given $m = 2\lfloor 2\log(p) \rfloor$, we consider the measures ν_0 and ν_1 as in Lemma 19. Given any measurable event A , we define μ_0 and μ_1 by

$$\mu_0(A) = (1 - \frac{\bar{k}_0}{p})\delta_0[M.A] + \frac{\bar{k}_1}{p}\nu_0[M.A], \quad \mu_1(A) = (1 - \frac{\bar{k}_1}{p})\delta_0[M.A] + \frac{\bar{k}_1}{p}\nu_1[M.A]. \quad (77)$$

Here, $M.A$ stands for $\{Mx : x \in A\}$ and δ_0 is the Dirac measure at 0. In view of this definition, the first m moments of μ_0 and μ_1 are matching.

Step 2. Properties of the priors. We consider the prior measures $\mu_0^{\otimes p}$ and $\mu_1^{\otimes p}$. In view of Lemma 14, we need to show that $\mu_0^{\otimes p}$ is concentrated on $\mathbb{B}_0[k_0]$ and that $\mu_1^{\otimes p}$ is concentrated on $\mathbb{B}_0[k_0, \Delta, \rho]$ for some large ρ .

Under $\mu_0^{\otimes p}$, $\|\theta\|_0$ follows a binomial distribution with parameter $(p, (k_0 - \Delta/2)/p)$. By Chebyshev's inequality,

$$\mu_0^{\otimes p}[\|\theta\|_0 > k_0] \leq 4 \frac{k_0}{\Delta^2} \leq 4p^{-1/8}, \quad (78)$$

since $\Delta \geq p^{1/4} \vee k_0^{2/3}$. Similarly,

$$\mu_0^{\otimes p}[\|\theta\|_0 \in (k_0 + \Delta/4, k_0 + 3\Delta/4)] \geq 1 - \frac{32k_0}{\Delta^2} \geq 1 - 32p^{-1/8}. \quad (79)$$

Under the event $\|\theta\|_0 \in (k_0 + \Delta/4, k_0 + 3\Delta/4)$, the corresponding parameter θ satisfies

$$d_2^2(\theta, \mathbb{B}_0[k_0]) \geq \frac{\Delta}{4} a_m^2 M^2 = \frac{c^2}{4} \Delta \frac{\log(p)}{n} \tanh^2 \left[\frac{1}{m} \arg \cosh \left[1 + \frac{2\bar{k}_0}{\Delta} \right] \right].$$

Since $\arg \cosh[1 + 2\bar{k}_0/\Delta] \leq \arg \cosh[1 + 2p] \leq 4\log(p)$ for $p \geq 2$ and since $\tanh(t) \geq 0.4t$ for any $t \in (0, 1)$, we deduce that

$$d_2^2(\theta, \mathbb{B}_0[k_0]) \geq c' \frac{\Delta}{n \log(p)} \arg \cosh^2 \left[1 + \frac{2\bar{k}_0}{\Delta} \right] \geq c' \frac{\Delta}{n \log(p)} \log^2 \left[1 + \frac{k_0}{\Delta} \right]. \quad (80)$$

$\arg \cosh(x) \geq \log(x)$. As a consequence, with probability $\mu_1^{\otimes p}$ larger than $1 - 32p^{-1/8}$, θ belongs to $\mathbb{B}_0[k_0, \Delta, \rho]$ with $\rho^2 = c' \frac{\Delta}{n \log(p)} \log^2 \left[1 + \frac{k_0}{\Delta} \right]$. To apply Lemma 14, it remains to bound the total variation distance between

$$\mathbf{P}_0 = \int \mathbb{P}_\theta \mu_0^{\otimes p}(d\theta) \quad \text{and} \quad \mathbf{P}_1 = \int \mathbb{P}_\theta \mu_1^{\otimes p}(d\theta).$$

Step 3. Control of $\|\mathbf{P}_0 - \mathbf{P}_1\|_{TV}$. For $j = 0, \dots, p$, define the distribution $\mathbf{P}_0^{(j)} = \int \mathbb{P}_\theta \mu_1^{\otimes j} \otimes \mu_0^{\otimes p-j}(d\theta)$ with $\mathbf{P}_0^{(0)} = \mathbf{P}_0$ and $\mathbf{P}_0^{(p)} = \mathbf{P}_1$. By triangular inequality, one has

$$\|\mathbf{P}_0 - \mathbf{P}_1\|_{TV} \leq \sum_{j=0}^p \|\mathbf{P}_0^{(j)} - \mathbf{P}_0^{(j+1)}\|_{TV}. \quad (81)$$

This upper bound greatly simplifies the following computations as the distributions $\mathbf{P}_0^{(j)}$ and $\mathbf{P}_0^{(j+1)}$ only differ by one coordinate. Unfortunately, we conjecture that our minimax lower bound in Theorem 1 is suboptimal in the regime where k_0 is close to \sqrt{p} precisely because of the upper bound (81). In the arguably simpler Gaussian sequence model [17], we have directly computed the χ^2 distances between the corresponding distributions \mathbf{P}_0 and \mathbf{P}_1 to obtain the sharp separation distance in all regimes. If we use instead the decomposition (81) for the Gaussian sequence model, this leads to a suboptimal lower bound for k_0 close to \sqrt{p} . To close this gap in the linear regression model, one would therefore need to directly handle the χ^2 distance between \mathbf{P}_0 and \mathbf{P}_1 but we were not able to do it.

In the following, we shall bound independently each of these p distances $\|\mathbf{P}_0^{(j)} - \mathbf{P}_0^{(j+1)}\|_{TV}$. Interestingly, $\mathbf{P}_0^{(j)}$ and $\mathbf{P}_0^{(j+1)}$ only differ by the distribution of the $j+1$ -th coordinate of θ . The general idea is to condition with respect to all the coordinates except the $j+1$ -th one so that we consider a linear regression model with only one covariate.

Let us write $g_0^{(j)}(Y|\mathbf{X})$ the conditional density of Y given \mathbf{X} under $\mathbf{P}_0^{(j)}$.

$$g_0^{(j)}(Y|\mathbf{X}) = (2\pi)^{-n/2} \int \exp\left(-\frac{\|Y - \mathbf{X}\theta\|_2^2}{2}\right) \mu_1^{\otimes j} \otimes \mu_0^{\otimes p-j}(d\theta).$$

Writing down $\mathbb{E}_{\mathbf{X}}$ the expectation with respect to \mathbf{X} , we have

$$\begin{aligned} 2\|\mathbf{P}_0^{(j)} - \mathbf{P}_0^{(j+1)}\|_{TV} &= \mathbb{E}_{\mathbf{X}} \left[\int \left| g_0^{(j)}(Y|\mathbf{X}) - g_0^{(j+1)}(Y|\mathbf{X}) \right| dY \right] \\ &= (2\pi)^{-n/2} \mathbb{E}_{\mathbf{X}} \left[\int \left| \int e^{-\|Y - \mathbf{X}\theta\|_2^2/2} \mu_0^{\otimes j} \otimes [\mu_0 - \mu_1] \otimes \mu_1^{\otimes p-j-1}(d\theta) \right| dY \right] \\ &= (2\pi)^{-n/2} \mathbb{E}_{\mathbf{X}} \left[\int \left| \int e^{-\|Y - \mathbf{X}\theta\|_2^2/2} \mu_0^{\otimes j} \otimes \mu_1^{\otimes p-j-1} \otimes [\mu_0 - \mu_1](d\theta) \right| dY \right], \end{aligned}$$

by permutation invariance. We call A_j this last quantity.

Given a p -dimensional vector θ , let $\theta^{(-p)}$ be such that $\theta_p^{(-p)} = 0$ and $\theta_j^{(-p)} = \theta_j$ for all $j < p$. Write $\boldsymbol{\mu}_j = \mu_0^{\otimes j} \otimes \mu_1^{\otimes p-j-1}$ and $\mu_\Delta = \mu_0 - \mu_1$.

$$(2\pi)^{n/2} A_j = \mathbb{E}_{\mathbf{X}} \left[\int \left| \int e^{-\|Y - \mathbf{X}\theta^{(-p)}\|_2^2/2} \left[\int e^{\omega_p \theta_p + \xi_p \theta_p^2} \mu_\Delta(d\theta_p) \right] \boldsymbol{\mu}_j(d\theta^{(-p)}) \right| dY \right],$$

where the quantities ω_p and ξ_p are defined by

$$\omega_p = (Y - \mathbf{X}\theta^{(-p)})^T \mathbf{X}_{\cdot,p}; \quad \xi_p = \sum_i \mathbf{X}_{i,p}^2.$$

Let Ω be the event such that $\Omega = \{|\omega_p| \leq 5\sqrt{n \log(p)}, \xi_p \leq 2n\}$.

Fix any $\theta \in \mathbb{R}^p$ such that $\|\theta\|_\infty \leq M$. Then, under \mathbb{P}_θ , ξ_p follows a χ^2 distribution with n degrees of freedom. As a consequence of deviation inequalities for χ^2 distributions (Lemma 3), its probability to be larger than $2n$ is smaller than $e^{-n/16}$. Besides, conditionally to \mathbf{X} , ω_p follows a

normal distribution with mean $\theta_p \|\mathbf{X}_p\|_2^2$ and variance $\|\mathbf{X}_p\|_2^2$. As a consequence, under the event $\{\xi_p \leq 2n\}$, the probability that $|\omega_p| \geq 2Mn + 2\sqrt{2n \log(p)}$ is smaller than $1/p^2$. In view of the definition (73) of M and by taking the constant c in that definition small enough, we conclude that

$$\mathbb{P}_\theta(\Omega) \geq 1 - 1/p^2 - e^{-n/16}, \quad (82)$$

for all θ such that $\|\theta\|_\infty \leq M$. We set

$$A_{j,\Omega} = (2\pi)^{-n/2} \mathbb{E}_{\mathbf{X}} \left[\int \left| \int e^{-\|Y - \mathbf{X}\theta^{(-p)}\|_2^2/2} \left[\int \mathbf{1}_\Omega e^{\omega_p \theta_p + \xi_p \theta_p^2} \mu_\Delta(d\theta_p) \right] \boldsymbol{\mu}_j(d\theta^{(-p)}) \right| dY \right],$$

It follows from this definition and from Equation (82) that

$$\begin{aligned} A_j &\leq A_{j,\Omega} + (2\pi)^{-n/2} \mathbb{E}_{\mathbf{X}} \left[\int \left| \int \mathbf{1}_{\bar{\Omega}} e^{-\|Y - \mathbf{X}\theta\|_2^2/2} \boldsymbol{\mu}_j(d\theta^{(-p)}) \mu_\Delta(d\theta_p) \right| dY \right] \\ &\leq A_{j,\Omega} + \sup_{\theta \in [-M, M]^p} \mathbb{P}_\theta[\bar{\Omega}] \leq A_{j,\Omega} + \frac{1}{p^2} + e^{-n/16}. \end{aligned} \quad (83)$$

In order to work out the term $A_{j,\Omega}$, we rely on the power expansion of $e^{\omega_p \theta_p + \xi_p \theta_p^2}$ together with the nullity of the m first moments of μ_Δ .

$$\begin{aligned} \left| \int \mathbf{1}_\Omega e^{\omega_p \theta_p + \xi_p \theta_p^2} \mu_\Delta(d\theta_p) \right| &= \left| \sum_{d=m/2+1}^{\infty} \int \mathbf{1}_\Omega \frac{(\omega_p \theta_p + \xi_p \theta_p^2/2)^d}{d!} \mu_\Delta(d\theta_p) \right| \\ &\leq \frac{\bar{k}_1}{p} \sum_{d=m/2+1}^{\infty} \frac{(5M\sqrt{n \log(p)} + nM^2)^d}{d!} \\ &\leq \frac{\bar{k}_1}{p} \sum_{d=m/2+1}^{\infty} \left[\frac{2e(5M\sqrt{n \log(p)} + nM^2)}{m} \right]^d \leq \frac{\bar{k}_1}{p2^{m/2}}, \end{aligned}$$

since, by (73), $m \geq 10e(5M\sqrt{n \log(p)} + nM^2)$ if we fix $c \leq 1/(10e)$. Plugging this bound into the definition of $A_{j,\Omega}$, we obtain

$$A_{j,\Omega} \leq \frac{\bar{k}_1}{p2^m} (2\pi)^{-n/2} \mathbb{E}_{\mathbf{X}} \left[\int \int e^{-\|Y - \mathbf{X}\theta^{(-p)}\|_2^2/2} \boldsymbol{\mu}_j(d\theta^{(-p)}) dY \right] \leq \frac{\bar{k}_1}{p2^m} \leq \frac{1}{p^2}, \quad (84)$$

by definition (73) of m . Together with (83), this implies that $A_j \leq 2/p^2 + e^{-n/16}$. Then, we use the definition of A_j and (81) to conclude that

$$\|\mathbf{P}_0 - \mathbf{P}_1\|_{TV} \leq \frac{2}{p} + pe^{-n/16},$$

which is smaller than $1/4$ for p large enough since the assumptions of Lemma 18 enforce that $p^{1/4} \leq n$. In view of the above bound, (78), (79), and (79), we are in position to apply Lemma 14. Thus, for p large enough, we conclude that

$$\rho_\gamma^{*2}[k_0, \Delta] \geq c \frac{\Delta}{n \log(p)} \log^2 \left[1 + \frac{k_0}{\Delta} \right].$$

□

6.3 Proof of Proposition 6

Since $\rho_{g,\gamma}^*[k_0, \Delta] \geq \rho_\gamma^*[k_0, \Delta]$, the second part of (21) comes from Theorem 1. Turning to the first part of (21), we have already pointed out in the proof of Lemma 17 that it is proved in [45] and [46] that, for all (k, n, p) one has

$$\rho_{g,\gamma}^{*2}[0, k] \geq c \begin{cases} \frac{k}{2n} \log \left(1 + \frac{\sqrt{p}}{k} \right) \\ -1 + \left(\frac{p}{2ek} \right)^{k/n} (4k)^{-2/n} & \text{if } k \log \left(\frac{\sqrt{p}}{e^{3/2}k} \right) \geq n. \end{cases}$$

These two bounds imply that, for p large enough and for all $\Delta \leq \sqrt{p}/e^3$, one has

$$\rho_{g,\gamma}^{*2}[0, \Delta] \geq c \frac{\Delta}{n} \log \left(1 + \frac{\sqrt{p}}{\Delta} \right) \exp \left[c' \frac{\Delta}{n} \log \left(1 + \frac{\sqrt{p}}{\Delta} \right) \right].$$

Since $\rho_{g,\gamma}^*[0, \Delta]$ is nondecreasing with respect to Δ (Lemma 15), the above bound is also valid for all $\Delta \leq p$ at the price of worse constants. Finally, we apply Lemma 15 together with the assumption $k_0 \leq n \leq p/2$ to obtain the first part of (21).

Acknowledgements.

The work of A. Carpentier is partially supported by the Deutsche Forschungsgemeinschaft (DFG) Emmy Noether grant MuSyAD (CA 1488/1-1), by the DFG - 314838170, GRK 2297 MathCoRe, by the DFG GRK 2433 DAEDALUS, by the DFG CRC 1294 'Data Assimilation', Project A03, and by the UFA-DFH through the French-German Doktorandenkolleg CDFA 01-18. The authors thank anonymous reviewers for their helpful suggestions that improved the manuscript. The authors are also grateful to Alexandre Tsybakov and Cun-Hui Zhang for bringing to our knowledge some recent work on MCP.

References

- [1] E. Arias-Castro, E. Candes, and Y. Plan. Global Testing under Sparse Alternatives: ANOVA, Multiple Comparisons and the Higher Criticism. *Annals of Statistics*, 39:2533–2556, 2011.
- [2] Yannick Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606, 2002.
- [3] Yannick Baraud, Sylvie Huet, and Béatrice Laurent. Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *Annals of statistics*, pages 214–257, 2005.
- [4] A. Belloni, V. Chernozhukov, and L. Wang. Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.
- [5] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.
- [6] Jelena Bradic, Jianqing Fan, and Yinchu Zhu. Testability of high-dimensional linear models with non-sparse structures. *arXiv preprint arXiv:1802.09117*, 2018.
- [7] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.
- [8] T Tony Cai and Zijian Guo. Accuracy assessment for high-dimensional linear regression. *arXiv preprint arXiv:1603.03474*, 2016.

- [9] T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.
- [10] T. Tony Cai and Jiashun Jin. Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Statist.*, 38(1):100–145, 2010.
- [11] T. Tony Cai, Jiashun Jin, and Mark G. Low. Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, 35(6):2421–2449, 2007.
- [12] T Tony Cai and Mark G Low. An adaptation theory for nonparametric confidence intervals. *Annals of statistics*, pages 1805–1840, 2004.
- [13] T Tony Cai and Mark G Low. Adaptive confidence balls. *The Annals of Statistics*, 34(1):202–228, 2006.
- [14] T. Tony Cai and Mark G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.*, 39(2):1012–1041, 2011.
- [15] Alexandra Carpentier. Testing the regularity of a smooth signal. *Bernoulli*, 21(1):465–488, 2015.
- [16] Alexandra Carpentier, Olivier Collier, Laëtitia Comminges, Alexandre B Tsybakov, and Yuhao Wang. Minimax rate of testing in sparse linear regression. *arXiv preprint arXiv:1804.06494*, 2018.
- [17] Alexandra Carpentier and Nicolas Verzelen. Adaptive estimation of the sparsity in the Gaussian vector model. *arXiv preprint arXiv:1703.00167*, 2017.
- [18] Olivier Collier, Laëtitia Comminges, and Alexandre B Tsybakov. Minimax estimation of linear and quadratic functionals on sparsity classes. *arXiv preprint arXiv:1502.00665*, 2015.
- [19] Olivier Collier, Laëtitia Comminges, Alexandre B Tsybakov, and Nicolas Verzelen. Optimal adaptive estimation of linear functionals under sparsity. *arXiv preprint arXiv:1611.09744*, 2016.
- [20] Laëtitia Comminges and Arnak S Dalalyan. Minimax testing of a composite null hypothesis defined via a quadratic functional in the model of regression. *Electronic journal of statistics*, 7:146–190, 2013.
- [21] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*, pages 317–366. North-Holland, Amsterdam, 2001.
- [22] Lee H. Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284, 2014.
- [23] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 2004.
- [24] Long Feng and Cun-Hui Zhang. Sorted concave penalized regression. *arXiv preprint arXiv:1712.09941*, 2017.
- [25] Ghislaine Gayraud and Christophe Pouet. Adaptive minimax testing in the discrete regression scheme. *Probability theory and related fields*, 133(4):531–558, 2005.
- [26] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2015.
- [27] Marc Hoffmann and Richard Nickl. On adaptive inference and confidence bands. *Ann. Statist.*, 39(5):2383–2409, 2011.
- [28] Y. Ingster, A. Tsybakov, and N. Verzelen. Detection boundary in sparse regression. *Electronic Journal of Statistics*, 4:1476–1526, 2010.
- [29] Yu. I. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2003.
- [30] Adel Javanmard and Jason D Lee. A Flexible Framework for Hypothesis Testing in High-

- dimensions. *arXiv preprint arXiv:1704.07971*, 2017.
- [31] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.
 - [32] Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.
 - [33] Jiashun Jin. Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(3):461–493, 2008.
 - [34] Jiashun Jin and T. Tony Cai. Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.*, 102(478):495–506, 2007.
 - [35] Anatoli Juditsky and Arkadi Nemirovski. On nonparametric tests of positivity/monotonicity/convexity. *Ann. Statist.*, 30(2):498–527, 2002.
 - [36] V. Koltchinski, K. Lounici, and A. Tsybakov. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011.
 - [37] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.
 - [38] Oleg Lepski, Arkady Nemirovski, and Vladimir Spokoiny. On estimation of the L r norm of a regression function. *Probability theory and related fields*, 113(2):221–253, 1999.
 - [39] Richard Nickl and Sara van de Geer. Confidence sets in sparse regression. *Ann. Statist.*, 41(6):2852–2876, 2013.
 - [40] G. Raskutti, M. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.
 - [41] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians. Volume III*, pages 1576–1602. Hindustan Book Agency, New Delhi, 2010.
 - [42] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.
 - [43] Alexandre B Tsybakov. Introduction to nonparametric estimation. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats, 2009.
 - [44] Sara Van de Geer, Peter Bühlmann, Ya’acov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.
 - [45] N. Verzelen and F. Villers. Goodness-of-fit tests for high-dimensional Gaussian linear models. *Annals of Statistics*, 38:704–752, 2010.
 - [46] Nicolas Verzelen. Minimax risks for sparse regressions: ultra-high dimensional phenomenons. *Electron. J. Stat.*, 6:38–90, 2012.
 - [47] Nicolas Verzelen and Elisabeth Gassiat. Adaptive estimation of High-Dimensional Signal-to-Noise Ratios. *arXiv preprint arXiv:1602.08006*, 2016.
 - [48] Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, 42(6):2164, 2014.
 - [49] Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.
 - [50] Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.
 - [51] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.
 - [52] Yinchu Zhu and Jelena Bradic. A projection pursuit framework for testing general high-dimensional hypothesis. *arXiv preprint arXiv:1705.01024*, 2017.

- [53] Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, pages 1–18, 2018.
- [54] Yinchu Zhu and Jelena Bradic. Significance testing in non-sparse high-dimensional linear models. *Electronic Journal of Statistics*, 12(2):3312–3364, 2018.